Optimal, Stochastic, Non-smooth, Non-convex Optimization through Online-to-Non-convex Conversion

Ashok Cutkosky [1] **Harsh Mehta** [2] **Francesco Orabona** [1]

Abstract

We present new algorithms for optimizing non-smooth, non-convex stochastic objectives based on a novel analysis technique. This improves the current best-known complexity for finding a (δ, ϵ)-stationary point from $O(\epsilon^{-4}\delta^{-1})$ stochastic gradient queries to $O(\epsilon^{-3}\delta^{-1})$, which we also show to be optimal. Our primary technique is a reduction from non-smooth non-convex optimization to *online learning*, after which our results follow from standard regret bounds in online learning. For *deterministic and second-order smooth* objectives, applying more advanced optimistic online learning techniques enables a new complexity of $O(\epsilon^{-1.5}\delta^{-0.5})$. Our improved non-smooth analysis also immediately recovers all optimal or best-known results for finding ϵ stationary points of smooth or second-order smooth objectives in both stochastic and deterministic settings.

1. Introduction

Algorithms for non-convex optimization are one of the most important tools in modern machine learning, as training neural networks requires optimizing a non-convex objective. Given the abundance of data in many domains, the time to train a neural network is the current bottleneck to having bigger and more powerful machine learning models. Motivated by this need, the past few years have seen an explosion of research focused on understanding non-convex optimization (Ghadimi & Lan, 2013; Carmon et al., 2017; Arjevani et al., 2019; 2020; Carmon et al., 2019; Fang et al., 2018). Despite significant progress, key issues remain unaddressed.

In this paper, we work to minimize a potentially non-convex objective $F : \mathbb{R}^d \to \mathbb{R}$ which we only accesss in some stochastic or "noisy" manner. As motivation, consider $F(\mathbf{x}) \triangleq \mathbb{E}_{\mathbf{z}}[f(\mathbf{x}, \mathbf{z})]$, where $\mathbf{x}$ can represent the model weights, $\mathbf{z}$ a minibatch of i.i.d. examples, and f the loss of a model with parameters $\mathbf{x}$ on the minibatch $\mathbf{z}$. In keeping with standard empirical practice, we will restrict ourselves to first order algorithms (gradient-based optimization).

The vast majority of prior analyses of non-convex optimization algorithms impose various smoothness conditions on the objective (Ghadimi & Lan, 2013; Carmon et al., 2017; Allen-Zhu, 2018; Tripuraneni et al., 2018; Fang et al., 2018; Zhou et al., 2018; Fang et al., 2019; Cutkosky & Orabona, 2019; Li & Orabona, 2019; Cutkosky & Mehta, 2020; Zhang et al., 2020a; Karimireddy et al., 2020; Levy et al., 2021; Faw et al., 2022; Liu et al., 2022). One motivation for smoothness assumptions is that they allow for a convenient surrogate for global minimization: rather than finding a global minimum of a neural network's loss surface (which may be intractable), we can hope to find an ϵ-stationary point, i.e., a point $\mathbf{x}$ such that $\|\nabla F(\mathbf{x})\| \leq \epsilon$. By now, the fundamental limits on first order smooth non-convex optimization are well understood: Stochastic Gradient Descent (SGD) will find an ϵ-stationary point in $O(\epsilon^{-4})$ iterations, which is the optimal rate (Arjevani et al., 2019). Moreover, if F happens to be second-order smooth, SGD requires only $O(\epsilon^{-3.5})$ iterations, which is also optimal (Fang et al., 2019; Arjevani et al., 2020). These optimality results motivate the popularity of SGD and its variants in practice (Kingma & Ba, 2014; Loshchilov & Hutter, 2016; 2018; Goyal et al., 2017; You et al., 2019).

Unfortunately, many standard neural network architectures are non-smooth (e.g., architectures incorporating ReLUs or max-pools cannot be smooth). As a result, these analyses can only provide intuition about what might occur when an algorithm is deployed in practice: the theorems themselves do not apply (see Patel & Berahas (2022) for examples of failure of SGD in non-smooth settings, or Li et al. (2021) for futher discussion of assumptions). Despite the obvious need for non-smooth analyses, recent results suggest that even approaching a neighborhood of a stationary point may be impossible for non-smooth objectives (Kornowski & Shamir, 2022b). Nevertheless, optimization clearly *is* possible in practice, which suggests that we may need to rethink our

[1]Department of Electrical & Computer Engineering, Boston University, Boston, MA, USA [2]Google Research, Mountain View, CA, USA. Correspondence to: Ashok Cutkosky <ashok@cutkosky.com>, Harsh Mehta <harshm@google.com>, Francesco Orabona <francesco@orabona.com>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

assumptions and goals in order to understand non-smooth optimization.

Fortunately, Zhang et al. (2020b) recently considered an alternative definition of stationarity that *is* tractable even for non-smooth objectives and which has attracted much interest (Davis et al., 2021; Tian et al., 2022; Kornowski & Shamir, 2022a; Tian & So, 2022; Jordan et al., 2022). Roughly speaking, instead of $\|\nabla F(\mathbf{x})\| \leq \epsilon$, we ask that there is a random variable $\mathbf{y}$ supported in a ball of radius δ about $\mathbf{x}$ such that $\|\mathbb{E}[\nabla F(\mathbf{y})]\| \leq \epsilon$. We call such an $\mathbf{x}$ an (δ, ϵ)-stationary point, so that the previous definition ($\|\nabla F(\mathbf{x})\| \leq \epsilon$) is a $(0, \epsilon)$-stationary point. The current best-known complexity for identifying an (δ, ϵ) stationary point is $O(\epsilon^{-4}\delta^{-1})$ stochastic gradient evaluations.

In this paper, we significantly improve this result: we can identify an (δ, ϵ)-stationary point with only $O(\epsilon^{-3}\delta^{-1})$ stochastic gradient evaluations. Moreover, we also show that this rate is optimal. Our primary technique is a novel *online-to-non-convex conversion*: a connection between non-convex stochastic optimization and *online learning*, which is a classical field of learning theory that already has a deep literature (Cesa-Bianchi & Lugosi, 2006; Hazan, 2019; Orabona, 2019). In particular, we show that an online learning algorithm that provides a shifting regret bound can be used to decide the update step, when fed with linear losses constructed using the stochastic gradients of the function F. By establishing this connection, we open new avenues for algorithm design in non-convex optimization and also motivate new research directions in online learning.

In sum, we make the following contributions:

- A reduction from non-convex non-smooth stochastic optimization to online learning: better online learning algorithms result in faster non-convex optimization. Applying this reduction to standard online learning algorithms allows us to identify an (δ, ϵ) stationary point in $O(\epsilon^{-3}\delta^{-1})$ stochastic gradient evaluations. The previous best-known rate in this setting was $O(\epsilon^{-4}\delta^{-1})$.
- We show that the $O(\epsilon^{-3}\delta^{-1})$ rate is optimal for all δ, ϵ such that $\epsilon \leq O(\delta)$.

Additionally, we prove important corollaries for smooth F:

- The $O(\epsilon^{-3}\delta^{-1})$ complexity implies the optimal $O(\epsilon^{-4})$ and $O(\epsilon^{-3.5})$ respective complexities for finding $(0, \epsilon)$-stationary points of smooth or second-order smooth objectives.
- For deterministic and second-order smooth objectives, we obtain a rate of $O(\epsilon^{-3/2}\delta^{-1/2})$, which implies the best-known $O(\epsilon^{-7/4})$ complexity for finding $(0, \epsilon)$-stationary points.

1.1. Related Work

In addition to the papers discussed in the introduction, here we discuss further related work.

In this paper we build on top of the definition of (δ, ϵ)-stationary points proposed by (Zhang et al., 2020b). There, they prove a complexity rate of $O(\epsilon^{-4}\delta^{-1})$ for stochastic Lipschitz functions, which we improve to $O(\epsilon^{-3}\delta^{-1})$ and prove the optimality of this result.

The idea to reduce machine learning to online learning was pioneered by Cesa-Bianchi et al. (2004) with the online-to-batch conversion. There is also previous work exploring the possibility of transforming non-convex problems into online learning ones. Ghai et al. (2022) provides some conditions under which online gradient descent on non-convex losses is equivalent to a convex online mirror descent. Hazan et al. (2017) defines a notion of regret which can be used to find approximate stationary points of smooth objectives. Zhuang et al. (2019) transform the problem of tuning of learning rates in stochastic non-convex optimization into an online learning problem. Our proposed approach differs from all the ones above in applying to non-smooth objectives. Moreover, as discusses in the next section, we employ online learning algorithms with *shifting* regret (Herbster & Warmuth, 1998) to generate the *updates* (i.e. the differences between successive iterates), rather than the iterates themselves.

2. Definitions and Setup

Here, we formally introduce our setting and notation. We are interested in optimizing real-valued functions $F : \mathcal{H} \to \mathbb{R}$ where $\mathcal{H}$ is a real Hilbert space (e.g., usually $\mathcal{H} = \mathbb{R}^d$). We assume $F^\star \triangleq \inf_{\mathbf{x}} F(\mathbf{x}) > -\infty$. We assume that F is differentiable, but we do *not* assume that F is smooth. All norms $\|\cdot\|$ are the Hilbert space norm (i.e., the 2-norm) unless otherwise specified. As mentioned in the introduction, the motivating example to keep in mind in our development is the case $F(\mathbf{x}) = \mathbb{E}_{\mathbf{z}}[f(\mathbf{x}, \mathbf{z})]$.

Our algorithms access information about F through a *stochastic gradient oracle* $\textsc{Grad} : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}$. Given a point $\mathbf{x}$ in $\mathcal{H}$, the oracle will sample an i.i.d. random variable $\mathbf{z} \in \mathcal{Z}$ and return $\textsc{Grad}(\mathbf{x}, \mathbf{z}) \in \mathcal{H}$ such that $\mathbb{E}[\textsc{Grad}(\mathbf{x}, \mathbf{z})] = \nabla F(\mathbf{x})$ and $\mathrm{Var}(\textsc{Grad}(\mathbf{x}, \mathbf{z})) \leq \sigma^2$.

In the following, we only consider functions satisfying the following mild regularity condition.

Definition 1. *We define a differentiable function $F : \mathcal{H} \to \mathbb{R}$ to be **well-behaved** if for all $\mathbf{x}, \mathbf{y} \in \mathcal{H}$, it holds that*

$$F(\mathbf{y}) - F(\mathbf{x}) = \int_0^1 \langle \nabla F(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle \, \mathrm{d}t \, .$$

If F happens to be differentiable and locally Lipschitz, then

this assumption is simply the Fundamental Theorem of Calculus. Under this assumption, our results can be applied to improve the past results on non-smooth stochastic optimization. In fact, Proposition 2 (proof in Appendix A) below shows that for the wide class of functions that are locally Lipschitz (but possibly non-differentiable), applying an arbitrarily small perturbation to the function is sufficient to ensure both differentiability and well-behavedness. This result works via standard perturbation arguments similar to those used previously by Davis et al. (2021) (see also Bertsekas (1973); Duchi et al. (2012); Flaxman et al. (2005) for similar techniques in the convex setting). In practice we suspect that such perturbation arguments are unnecessary: intuitively an algorithm is unlikely to query a point of non-differentiability (see also Bianchi et al. (2022) for some formal evidence for this idea).

Proposition 2. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be locally Lipschitz with stochastic oracle* GRAD *such that $\mathbb{E}_{\mathbf{z}}[\text{GRAD}(\mathbf{x}, \mathbf{z})] = \nabla F(\mathbf{x})$ whenever F is differentiable. We have two cases:*

- *If F is differentiable everywhere, then F is well-behaved.*
- *If F is not differentiable everywhere, let $p > 0$ be an arbitrary number and let $\mathbf{u}$ be a random vector in $\mathbb{R}^d$ uniformly distributed on the unit ball. Define $\hat{F}(\mathbf{x}) \triangleq \mathbb{E}_{\mathbf{u}}[F(\mathbf{x} + p\mathbf{u})]$. Then, $\hat{F}$ is differentiable and well-behaved, and the oracle $\widehat{\text{GRAD}}(\mathbf{x}, (\mathbf{z}, \mathbf{u})) = \text{GRAD}(\mathbf{x}+p\mathbf{u}, \mathbf{z})$ is a stochastic gradient oracle for $\hat{F}$. Moreover, F is differentiable at $\mathbf{x}+p\mathbf{u}$ with probability 1 and if F is G-Lipschitz, then $|\hat{F}(\mathbf{x}) - F(\mathbf{x})| \leq pG$ for all $\mathbf{x}$.*

Remark 3. *We explicitly note that our results cover the case in which F is* directionally *differentiable and we have access to a stochastic* directional *gradient oracle, as considered by Zhang et al. (2020b). This is a less standard oracle* $\text{GRAD}(\mathbf{x}, \mathbf{v}, \mathbf{z})$ *that outputs $\mathbf{g}$ such that $\mathbb{E}[\langle \mathbf{g}, \mathbf{v}\rangle]$ is the* directional derivative *of F in the direction $\mathbf{v}$. This setting is subtly different (although a directional derivative oracle is a gradient oracle at all points for which F is continuously differentiable). In order to keep technical complications to a minimum, in the main text we consider the simpler stochastic gradient oracle discussed above. In Appendix H, we show that our results and techniques also apply using directional gradient oracles with only superficial modification.*

2.1. (δ, ϵ)-Stationary Points

Now, let us define our notion of (δ, ϵ)-stationary point. This definition is essentially the same as used in Zhang et al. (2020b); Davis et al. (2021); Tian et al. (2022). It is in fact mildly more stringent since we restrict to distributions of finite support and require an "unbiasedness" condition in order to make eventual connections to second-order smooth objectives easier.

Definition 4. *A point $\mathbf{x}$ is an (δ, ϵ)-stationary point of an almost-everywhere differentiable function F if there is a finite subset S of the ball of radius δ centered at $\mathbf{x}$ such that for $\mathbf{y}$ selected uniformly at random from S, $\mathbb{E}[\mathbf{y}] = \mathbf{x}$ and $\|\mathbb{E}[\nabla F(\mathbf{z})]\| \leq \epsilon$.*

As a counterpart to this definition, we also define:

Definition 5. *Given a point $\mathbf{x}$, a number $\delta > 0$ and a almost-everywhere differentiable function F, define*

$$\|\nabla F(\mathbf{x})\|_\delta \triangleq \inf_{S \subset B(\mathbf{x}, \delta), \frac{1}{|S|}\sum_{\mathbf{y}\in S}\mathbf{y} = \mathbf{x}} \left\| \frac{1}{|S|} \sum_{\mathbf{y}\in S} \nabla F(\mathbf{y}) \right\| .$$

Let's also state an immediate corollary of Proposition 2 that converts a guarantee on a randomized smoothed function to one on the original function. This result is also immediate from Theorem 3.1 of Lin et al. (2022).

Corollary 6. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be G-Lipschitz. For $\epsilon > 0$, let $p \leq \delta$ and let $\mathbf{u}$ be a random vector in $\mathbb{R}^d$ uniformly distributed on the unit ball. Define $\hat{F}(\mathbf{x}) \triangleq \mathbb{E}_{\mathbf{u}}[F(\mathbf{x}+p\mathbf{u})]$. If a point $\mathbf{x}$ satisfies $\|\nabla \hat{F}(\mathbf{x})\|_\delta \leq \epsilon$, then $\|\nabla F(\mathbf{x})\|_{2\delta} \leq \epsilon$.*

Our ultimate goal is to use N stochastic gradient evaluations of F to identify a point $\mathbf{x}$ with as small a value of $\mathbb{E}[\|\nabla F(\mathbf{x})\|_\delta]$ as possible. For the rest of this paper we will consider exclusively the case of well-behaved and differentiable objectives F. We focus our development on this conceptually simpler case in order to simplify the proofs as much as possible, however due to Proposition 2 and Corollary 6, our results will immediately extend from differentiable F to those F that are locally Lipschitz and for which $\text{GRAD}(\mathbf{x}, \mathbf{z})$ returns a unbiased estimate of $\nabla F(\mathbf{x})$ whenever F is differentiable at $\mathbf{x}$.

2.2. Online Learning

Here, we very briefly introduce the setting of online linear learning with shifting competitors, that will be the core of our online-to-non-convex conversion. We refer the interested reader to Cesa-Bianchi & Lugosi (2006); Hazan (2019); Orabona (2019) for a comprehensive introduction to online learning. In the online learning setting, the learning process goes on in rounds. In each round the algorithm outputs a point $\mathbf{\Delta}_t$ in a feasible set V, and then receives a linear loss function $\ell_t(\cdot) = \langle \mathbf{g}_t, \cdot\rangle$ and it pays $\ell_t(\mathbf{\Delta}_t)$. The goal of the algorithm is to minimize the *static regret* over T rounds, defined as the difference between its cumulative loss and the one of an arbitrary comparison vector $\mathbf{u} \in V$:

$$R_T(\mathbf{u}) \triangleq \sum_{t=1}^{T} \langle \mathbf{g}_t, \mathbf{\Delta}_t - \mathbf{u}\rangle .$$

With *no stochastic assumption*, it is possible to design online algorithms that guarantee that the regret is upper bounded by $O(\sqrt{T})$. In this work, we frequently make use of a more challenging objective: minimizing the K-*shifting regret*. This is the regret with respect to an arbitrary *sequence* of K vectors $\mathbf{u}^1, \dots, \mathbf{u}^K \in V$ that changes every T iterations:

$$R_T(\mathbf{u}^1, \dots, \mathbf{u}^K) \triangleq \sum_{k=1}^{K} \sum_{n=(k-1)T+1}^{kT} \langle \mathbf{g}_n, \boldsymbol{\Delta}_n - \mathbf{u}^k \rangle \ . \quad (1)$$

It should be intuitive that resetting the online algorithm every T iterations can achieve a shifting regret of $O(K\sqrt{T})$.

3. Online-to-Non-Convex Conversion

In this section, we explain the online-to-non-convex conversion. The core idea transforms the minimization of a non-convex and non-smooth function onto the problem of minimizing the shifting regret over linear losses. In particular, consider an optimization algorithm that updates a previous iterate $\mathbf{x}_{n-1}$ by moving in a direction $\boldsymbol{\Delta}_n$: $\mathbf{x}_n = \mathbf{x}_{n-1} + \boldsymbol{\Delta}_n$. For example, SGD sets $\boldsymbol{\Delta}_n = -\eta \mathbf{g}_{n-1} = -\eta \cdot \textsc{Grad}(\mathbf{x}_{n-1}, \mathbf{z}_{n-1})$ for a learning rate η. Instead, we let an online learning algorithm $\mathcal{A}$ decide the update direction $\boldsymbol{\Delta}_n$, using linear losses $\ell_n(\mathbf{x}) = \langle \mathbf{g}_n, \mathbf{x} \rangle$.

The motivation behind essentially all first order algorithms is that $F(\mathbf{x}_{n-1} + \boldsymbol{\Delta}_n) - F(\mathbf{x}_{n-1}) \approx \langle \mathbf{g}_n, \boldsymbol{\Delta}_n \rangle$. This suggests that $\boldsymbol{\Delta}_n$ should be chosen to minimize the inner product $\langle \mathbf{g}_n, \boldsymbol{\Delta}_n \rangle$. However, we are faced with two difficulties. The first difficulty is the the approximation error in the first-order expansion. The second is the fact that $\boldsymbol{\Delta}_n$ needs to be chosen before $\mathbf{g}_n$ is revealed, so that $\boldsymbol{\Delta}_n$ needs in some sense to "predict the future". Typical analysis of algorithms such as SGD use the remainder form of Taylor's theorem to address both difficulties simultaneously for smooth objectives, but in our non-smooth case this is not a valid approach. Instead, we tackle these difficulties independently. We overcome the first difficulty using the same randomized scaling trick employed by Zhang et al. (2020b): define $\mathbf{g}_n$ to be a gradient evaluated not at $\mathbf{x}_{n-1}$ or $\mathbf{x}_{n-1} + \boldsymbol{\Delta}_n$, but at a *random* point along the line segment connecting the two. Then for a well-behaved function we will have $F(\mathbf{x}_{n-1} + \boldsymbol{\Delta}_n) - F(\mathbf{x}_{n-1}) = \mathbb{E}[\langle \mathbf{g}_n, \boldsymbol{\Delta}_n \rangle]$. The second difficulty is where online learning shines: online learning algorithms are *specifically designed* to predict *completely arbitrary* sequences of vectors as accurately as possible.

The previous intuition is formalized in Algorithm 1 and the following result, which we will elaborate on in Theorem 8 before yielding our main result in Corollary 9.

Algorithm 1 Online-to-Non-Convex Conversion

Input: Initial point $\mathbf{x}_0$, $K \in \mathbb{N}$, $T \in \mathbb{N}$, online learning algorithm $\mathcal{A}$.
Set $M = K \cdot T$
for $n = 1 \dots M$ **do**
  Get $\boldsymbol{\Delta}_n$ from $\mathcal{A}$
  Set $\mathbf{x}_n = \mathbf{x}_{n-1} + \boldsymbol{\Delta}_n$
  Generate $s_n \in [0, 1]$ // usually uniformly random, see Theorem statements for precise settings.
  Set $\mathbf{w}_n = \mathbf{x}_{n-1} + s_n \boldsymbol{\Delta}_n$
  Sample random $\mathbf{z}_n$
  Generate gradient $\mathbf{g}_n = \textsc{Grad}(\mathbf{w}_n, \mathbf{z}_n)$
  Send $\mathbf{g}_n$ to $\mathcal{A}$ as gradient
end for
Set $\mathbf{w}_t^k = \mathbf{w}_{(k-1)T+t}$ for $k = 1, \dots, K$ and $t = 1, \dots, T$
Set $\overline{\mathbf{w}}^k = \frac{1}{T} \sum_{t=1}^{T} \mathbf{w}_t^k$ for $k = 1, \dots, K$
Return $\{\overline{\mathbf{w}}^1, \dots, \overline{\mathbf{w}}^K\}$

Theorem 7. *Suppose F is well-behaved. Define $\boldsymbol{\nabla}_n = \int_0^1 \nabla F(\mathbf{x}_{n-1} + s\boldsymbol{\Delta}_n) \, \mathrm{d}s$. Then, with the notation in Algorithm 1 and for any sequence of vectors $\mathbf{u}_1, \dots, \mathbf{u}_N$, we have the* equality*:*

$$F(\mathbf{x}_M) = F(\mathbf{x}_0) + \sum_{n=1}^{M} \langle \mathbf{g}_n, \boldsymbol{\Delta}_n - \mathbf{u}_n \rangle + \sum_{n=1}^{M} \langle \boldsymbol{\nabla}_n - \mathbf{g}_n, \boldsymbol{\Delta}_n \rangle + \sum_{n=1}^{M} \langle \mathbf{g}_n, \mathbf{u}_n \rangle \ .$$

Moreover, if we let s_n be independent random variables uniformly distributed in $[0, 1]$, then we have

$$\mathbb{E}[F(\mathbf{x}_M)] = F(\mathbf{x}_0) + \mathbb{E}\left[\sum_{n=1}^{M} \langle \mathbf{g}_n, \boldsymbol{\Delta}_n - \mathbf{u}_n \rangle\right] + \mathbb{E}\left[\sum_{n=1}^{M} \langle \mathbf{g}_n, \mathbf{u}_n \rangle\right] \ .$$

Proof. By the well-behaveness of F, we have

$$\begin{aligned} & F(\mathbf{x}_n) - F(\mathbf{x}_{n-1}) \\ &= \int_0^1 \langle \nabla F(\mathbf{x}_{n-1} + s(\mathbf{x}_n - \mathbf{x}_{n-1})), \mathbf{x}_n - \mathbf{x}_{n-1} \rangle \, \mathrm{d}s \\ &= \int_0^1 \langle \nabla F(\mathbf{x}_{n-1} + s\boldsymbol{\Delta}_n), \boldsymbol{\Delta}_n \rangle \, \mathrm{d}s \\ &= \langle \boldsymbol{\nabla}_n, \boldsymbol{\Delta}_n \rangle \\ &= \langle \mathbf{g}_n, \boldsymbol{\Delta}_n - \mathbf{u}_n \rangle + \langle \boldsymbol{\nabla}_n - \mathbf{g}_n, \boldsymbol{\Delta}_n \rangle + \langle \mathbf{g}_n, \mathbf{u}_n \rangle \ . \end{aligned}$$

Now, sum over n and telescope to obtain the stated bound.

For the second statement, simply observe that by definition we have $\mathbb{E}[\mathbf{g}_n] = \int_0^1 \nabla F(\mathbf{x}_{n-1} + s\boldsymbol{\Delta}_n) \, \mathrm{d}s = \boldsymbol{\nabla}_n$. □

3.1. Guarantees for Non-Smooth Non-Convex Functions

The primary value of Theorem 7 is that the term $\sum_{n=1}^M \langle \mathbf{g}_n, \boldsymbol{\Delta}_n - \mathbf{u}_n \rangle$ is exactly the regret of an online learning algorithm: lower regret clearly translates to a smaller bound on $F(\mathbf{x}_M)$. Next, by carefully choosing $\mathbf{u}_n$, we will be able to relate the term $\sum_{n=1}^M \langle \mathbf{g}_n, \mathbf{u}_n \rangle$ to the gradient averages that appear in the definition of (δ, ϵ)-stationarity. Formalizing these ideas, we have the following:

Theorem 8. *Assume F is well-behaved. With the notation in Algorithm 1, set s_n to be a random variable sampled uniformly from $[0,1]$. Set $T, K \in \mathbb{N}$ and $M = KT$. Define $\mathbf{u}^k = -D \frac{\sum_{t=1}^T \nabla F(\mathbf{w}_t^k)}{\|\sum_{t=1}^T \nabla F(\mathbf{w}_t^k)\|}$ for some $D > 0$ for $k = 1, \dots, K$. Finally, suppose Var$(\mathbf{g}_n) \leq \sigma^2$. Then:*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^K \left\| \frac{1}{T}\sum_{t=1}^T \nabla F(\mathbf{w}_t^k) \right\|\right] \leq \frac{F(\mathbf{x}_0) - F^\star}{DM} + \frac{\mathbb{E}[R_T(\mathbf{u}^1, \dots, \mathbf{u}^K)]}{DM} + \frac{\sigma}{\sqrt{T}} .$$

Proof. From Theorem 7, we have

$$\mathbb{E}[F(\mathbf{x}_M)] = F(\mathbf{x}_0) + \mathbb{E}[R_T(\mathbf{u}^1, \dots, \mathbf{u}^K)] + \mathbb{E}\left[\sum_{n=1}^M \langle \mathbf{g}_n, \mathbf{u}_n \rangle\right] .$$

Now, since $\mathbf{u}^k = -D \frac{\sum_{t=1}^T \nabla F(\mathbf{w}_t^k)}{\|\sum_{t=1}^T \nabla F(\mathbf{w}_t^k)\|}$, $\mathbb{E}[\mathbf{g}_n] = \mathbb{E}[\nabla F(\mathbf{w}_n)]$, and Var$(\mathbf{g}_n) \leq \sigma^2$, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{n=1}^M \langle \mathbf{g}_n, \mathbf{u}_n \rangle\right] &\leq \mathbb{E}\left[\sum_{k=1}^K \left\langle \sum_{t=1}^T \nabla F(\mathbf{w}_t^k), \mathbf{u}^k \right\rangle\right] \\
&\quad + \mathbb{E}\left[D \sum_{k=1}^K \left\| \sum_{t=1}^T (\nabla F(\mathbf{w}_t^k) - \mathbf{g}_{T(k-1)+t}) \right\|\right] \\
&\leq \mathbb{E}\left[\sum_{k=1}^K \left\langle \sum_{t=1}^T \nabla F(\mathbf{w}_t^k), \mathbf{u}^k \right\rangle\right] + D\sigma K\sqrt{T} \\
&= \mathbb{E}\left[-\sum_{k=1}^K DT \left\| \frac{1}{T}\sum_{t=1}^T \nabla F\left(\mathbf{w}_t^k\right) \right\|\right] + D\sigma K\sqrt{T} .
\end{aligned}$$

Putting this all together, we have

$$\begin{aligned}
F^\star \leq \mathbb{E}[F(\mathbf{x}_M)] &\leq F(\mathbf{x}_0) + \mathbb{E}[R_T(\mathbf{u}^1, \dots, \mathbf{u}^K)] \\
&\quad + \sigma DK\sqrt{T} - DT \sum_{k=1}^K \mathbb{E}\left[\left\| \frac{1}{T}\sum_{t=1}^T \nabla F\left(\mathbf{w}_t^k\right) \right\|\right] .
\end{aligned}$$

Dividing by $KDT = DM$ and reordering, we have the stated bound. □

We now instantiate Theorem 8 with the simplest online learning algorithm: online gradient descent (OGD) (Zinkevich, 2003). OGD takes input a radius D and a step size η and makes the update $\boldsymbol{\Delta}_{n+1} = \Pi_{\|\boldsymbol{\Delta}\| \leq D}[\boldsymbol{\Delta}_n - \eta \mathbf{g}_n]$ with $\boldsymbol{\Delta}_1 = 0$. The standard analysis shows that if $\mathbb{E}[\|\mathbf{g}_n\|^2] \leq G^2$ for all n, then with $\eta = \frac{D}{G\sqrt{T}}$, OGD will ensure[1] static regret $\mathbb{E}[R_T(\mathbf{u})] \leq DG\sqrt{T}$ for *any* $\mathbf{u}$ satisfying $\|\mathbf{u}\| \leq D$. Thus, by resetting the algorithm every T iterations, we achieve $\mathbb{E}[R_T(\mathbf{u}^1, \dots \mathbf{u}^K)] \leq KDG\sqrt{T}$. This powerful guarantee for all sequences is characteristic of online learning. We are now free to optimize the remaining parameters K and D to achieve our main result, presented in Corollary 9.

Corollary 9. *Suppose we have a budget of N gradient evaluations. Under the assumptions of Theorem 8, suppose in addition $\mathbb{E}[\|\mathbf{g}_n\|^2] \leq G^2$ and that $\mathcal{A}$ guarantees $\|\boldsymbol{\Delta}_n\| \leq D$ for some user-specified D for all n and ensures the worst-case K-shifting regret bound $\mathbb{E}[R_T(\mathbf{u}^1, \dots, \mathbf{u}^K)] \leq DGK\sqrt{T}$ for all $\|\mathbf{u}^k\| \leq D$ (e.g., as achieved by the OGD algorithm that is reset every T iterations). Let $\delta > 0$ be an arbitrary number. Set $D = \delta/T$, $T = \min(\lceil (\frac{GN\delta}{F(\mathbf{x}_0) - F^\star})^{2/3} \rceil, \frac{N}{2})$, and $K = \lfloor \frac{N}{T} \rfloor$. Then, for all k and t, we have $\|\overline{\mathbf{w}}^k - \mathbf{w}_t^k\| \leq \delta$.*

Moreover, we have the inequality

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^K \left\| \frac{1}{T}\sum_{t=1}^T \nabla F(\mathbf{w}_t^k) \right\|\right] \leq \frac{2(F(\mathbf{x}_0) - F^\star)}{\delta N} + \max\left(\frac{5G^{2/3}(F(\mathbf{x}_0) - F^\star)^{1/3}}{(N\delta)^{1/3}}, \frac{6G}{\sqrt{N}}\right),$$

which implies

$$\mathbb{E}\left[\frac{1}{K}\sum_{t=1}^K \|\nabla F(\overline{\mathbf{w}}^k)\|_\delta\right] \leq \frac{2(F(\mathbf{x}_0) - F^\star)}{\delta N} + \max\left(\frac{5G^{2/3}(F(\mathbf{x}_0) - F^\star)^{1/3}}{(N\delta)^{1/3}}, \frac{6G}{\sqrt{N}}\right) .$$

Before providing the proof, let us discuss the implications. Notice that if we select $\hat{\mathbf{w}}$ at random from $\{\overline{\mathbf{w}}^1, \dots, \overline{\mathbf{w}}^K\}$, then we clearly have $\mathbb{E}[\|\nabla F(\hat{\mathbf{w}})\|_\delta] = \mathbb{E}\left[\frac{1}{K}\sum_{t=1}^K \|\nabla F(\overline{\mathbf{w}}^k)\|_\delta\right]$. Therefore, the Corollary asserts that for a function F with $F(\mathbf{x}_0) - \inf F(\mathbf{x}) \leq \gamma$ with a stochastic gradient oracle whose second moment is bounded by G^2, a properly instantiated Algorithm 1 finds a (δ, ϵ) stationary point in $N = O(G\gamma\epsilon^{-3}\delta^{-1})$ gradient evaluations. In Section 7, we will provide a lower bound showing that this rate is optimal essentially whenever $\delta G^2 \geq \epsilon\gamma$. Together, the Corollary and the lower bound provide a nearly complete characterization of the complexity of finding (δ, ϵ)-stationary points in the stochastic setting.

[1] For completeness a proof of this statement is in Appendix B.

It is also interesting to note that the bound does not appear to improve if the gradients are deterministic. Specifically, in the assumptions for Corollary 9, we could try to relax $\mathbb{E}[\|\mathbf{g}_t\|^2] \le G$ to $\|\nabla F(\mathbf{w}_t)\| \le G$ and $\mathrm{Var}(\mathbf{g}_t) \le \sigma^2$ for some σ. We might then hope to improve the bound as $\sigma \to 0$ by taking advantage of the σ-dependency in Theorem 8. However, it turns out that the σ-dependency in Corollary 9 is dominated by a dependency on G coming from the regret bound of OGD. This highlights an interesting open question: is it actually possible to improve in the deterministic setting? It is conceivable that the answer is "no": in the non-smooth *convex* optimization setting, it is well-known that the optimal rates for stochastic and deterministic optimization are the same (see, e.g., Bubeck (2015) for proofs of both upper and lower bounds).

Remark 10. *We conjecture that by employing martingale concentration, the above can be extended to identify a* $(\delta, O(\frac{G^{2/3}(F(\mathbf{x}_0)-F^\star)^{1/3}}{(N\delta)^{1/3}}))$*-stationary point with high probability, although we do not establish such a result here.*

It is also interesting to explicitly write the update of the overall algorithm:

$$\begin{aligned}\mathbf{x}_n &= \mathbf{x}_{n-1} + \mathbf{\Delta}_n \\ \mathbf{g}_n &= \textsc{Grad}(\mathbf{x}_n + (s_n - 1)\mathbf{\Delta}_n, \mathbf{z}_n) \\ \mathbf{\Delta}_{n+1} &= \mathrm{clip}_D(\mathbf{\Delta}_n + \eta \mathbf{g}_n)\end{aligned}$$

where $\mathrm{clip}_D(\mathbf{x}) = \mathbf{x}\min(\frac{D}{\|\mathbf{x}\|}, 1)$. In words, the update is reminiscent of the SGD update with momentum and clipping. The primary different element is the fact that the stochastic gradient is taken on a slightly perturbed $\mathbf{x}_n$.

Proof of Corollary 9. Since $\mathcal{A}$ guarantees $\|\mathbf{\Delta}_n\| \le D$, for all $n < n' \le n + T - 1$, we have

$$\begin{aligned}\|\mathbf{w}_n - \mathbf{w}_{n'}\| &= \|\mathbf{x}_n - (1 - s_n)\mathbf{\Delta}_n - \mathbf{x}_{n'-1} + s_{n'}\mathbf{\Delta}_{n'}\| \\ &\le \left\|\sum_{i=n+1}^{n'-1} \mathbf{\Delta}_i\right\| + \|\mathbf{\Delta}_n\| + \|\mathbf{\Delta}_{n'}\| \\ &\le D((n'-1) - (n+1) + 1) + 2D \le DT\,.\end{aligned}$$

Therefore, we clearly have $\|\mathbf{w}_t^k - \overline{\mathbf{w}}^k\| \le DT = \delta$.

Note that from the choice of K and T we have $M = KT \ge N - T \ge N/2$. So, for the second fact, notice that $\mathrm{Var}(\mathbf{g}_n) \le \mathbb{E}[\|\mathbf{g}_t\|^2] \le G^2$ for all n. Thus, applying Theorem 8 in concert with the additional assumption $\mathbb{E}[R_T(\mathbf{u}^1, \dots, \mathbf{u}^K)] \le DGK\sqrt{T}$, we have:

$$\begin{aligned}&\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^K \left\|\frac{1}{T}\sum_{t=1}^T \nabla F(\mathbf{w}_t^k)\right\|\right] \\ &\le 2\frac{F(\mathbf{x}_0) - F^\star}{DN} + 2\frac{KDG\sqrt{T}}{DN} + \frac{G}{\sqrt{T}} \\ &\le \frac{2T(F(\mathbf{x}_0) - F^\star)}{\delta N} + \frac{3G}{\sqrt{T}} \\ &\le \max\left(\frac{5G^{2/3}(F(\mathbf{x}_0) - F^\star)^{1/3}}{(N\delta)^{1/3}}, \frac{6G}{\sqrt{N}}\right) \\ &\quad + \frac{2(F(\mathbf{x}_0) - F^\star)}{\delta N},\end{aligned}$$

where the last inequality is due to the choice of T.

Finally, observe that $\|\mathbf{w}_t^k - \overline{\mathbf{w}}^k\| \le \delta$ for all t and k, and also that $\overline{\mathbf{w}}^k = \frac{1}{T}\sum_{t=1}^T \mathbf{w}_t^k$. Therefore $S = \{\mathbf{w}_1^k, \dots, \mathbf{w}_T^k\}$ satisfies the conditions in the infimum in Definition 5 so that $\|\nabla F(\overline{\mathbf{w}}^k)\|_\delta \le \left\|\frac{1}{T}\sum_{t=1}^T \nabla F(\mathbf{w}_t^k)\right\|$. □

4. Bounds for the L_1 Norm

It is a well-known trick in the online learning literature that running a separate instance of an online learning algorithm on each coordinate of $\mathbf{\Delta}$ yields regret bounds with respect to L_1 norms of the linear costs (e.g., as in AdaGrad (Duchi et al., 2010; McMahan & Streeter, 2010)). For example, we can run the online gradient descent algorithm with a separate learning rate for each coordinate: $\mathbf{\Delta}_{n+1,i} = \Pi_{[-D_\infty, D_\infty]}[\mathbf{\Delta}_{n,i} - \eta_i \mathbf{g}_{n,i}]$. The regret of this procedure is simply the sum of the regrets of each of the individual algorithms. In particular, if $\mathbb{E}[\mathbf{g}_{n,i}^2] \le G_i^2$, then setting $\eta_i = \frac{D_\infty}{G_i\sqrt{T}}$ yields the regret bound $\mathbb{E}[R_T(\mathbf{u})] \le D_\infty\sqrt{T}\sum_{i=1}^N G_i$ for any $\mathbf{u}$ satisfying $\|\mathbf{u}\|_\infty \le D_\infty$. By employing such online algorithms with our online-to-non-convex conversion, we can obtain a guarantee on the L_1 norm of the gradients.

Definition 11. *A point* $\mathbf{x}$ *is a* (δ, ϵ)*-stationary point with respect to the* L_1 *norm of an almost-everywhere differentiable function* F *if there exists a finite subset* S *of the* L_∞ *ball of radius* δ *centered at* $\mathbf{x}$ *such that if* $\mathbf{y}$ *is selected uniformly at random from* S, $\mathbb{E}[\mathbf{y}] = \mathbf{x}$ *and* $\|\mathbb{E}[\nabla F(\mathbf{y})]\|_1 \le \epsilon$.

As a counterpart to this definition, we define:

Definition 12. *Given a point* $\mathbf{x}$*, a number* $\delta > 0$ *and an almost-everywhere differentiable function* F*, define*

$$\|\nabla F(\mathbf{x})\|_{1,\delta} \triangleq \inf_{S \subset B_\infty(\mathbf{x},\delta)|, \frac{1}{|S|}\sum_{\mathbf{y}\in S}\mathbf{y} = \mathbf{x}} \left\|\frac{1}{|S|}\sum_{\mathbf{y}\in S}\nabla F(\mathbf{y})\right\|_1.$$

We now can state a theorem similar to Corollary 9. Given that the proof is also very similar, we defer it to Appendix G.

Theorem 13. *Suppose we have a budget of N gradient evaluations. Assume $F : \mathbb{R}^d \to \mathbb{R}$ is well-behaved. With the notation in Algorithm 1, set s_n to be a random variable sampled uniformly from $[0, 1]$. Set $T, K \in \mathbb{N}$ and $M = KT$. Assume that $\mathbb{E}[g_{n,i}^2] \leq G_i^2$ for $i = 1, \dots, d$ for all n. Assume that $\mathcal{A}$ guarantees $\|\mathbf{\Delta}_n\|_\infty \leq D_\infty$ for some user-specified D_∞ for all n and ensures the K-shifting regret bound $\mathbb{E}[R_T(\mathbf{u}^1, \dots, \mathbf{u}^K)] \leq D_\infty K\sqrt{T}\sum_{i=1}^d G_i$ for all $\|\mathbf{u}^k\|_\infty \leq D_\infty$. Let $\delta > 0$ be an arbitrary number. Set $D_\infty = \delta/T$, $T = \min(\lceil(\frac{N\delta\sum_{i=1}^d G_i}{F(\mathbf{x}_0)-F^\star})^{2/3}\rceil, \frac{N}{2})$, and $K = \lfloor\frac{N}{T}\rfloor$. Then we have:*

$$\frac{1}{K}\sum_{t=1}^K \|\nabla F(\overline{\mathbf{w}}^k)\|_{1,\delta} \leq \frac{2(F(\mathbf{x}_0) - F^\star)}{\delta N} + \max\left(\frac{5(\sum_{i=1}^d G_i)^{2/3}(F(\mathbf{x}_0) - F^\star)^{1/3}}{(N\delta)^{1/3}}, \frac{6\sum_{i=1}^d G_i}{\sqrt{N}}\right).$$

Let's compare this result with Corollary 9. For a fair comparison, we set G_i and G such that $\sum_{i=1}^d G_i^2 = G^2$. Then, we can lower bound $\|\cdot\|_\delta$ with $\frac{1}{\sqrt{d}}\|\cdot\|_{1,\delta}$. Hence, under the assumption $\mathbb{E}[\|\mathbf{g}_n\|^2] = \sum_{i=1}^d \mathbb{E}[g_{n,i}^2] \leq \sum_{i=1}^d G_i^2 = G^2$, Corollary 9 implies $\frac{1}{K}\sum_{t=1}^K \|\nabla F(\overline{\mathbf{w}}^k)\|_{1,\delta} \leq O(\frac{G^{2/3}\sqrt{d}}{(N\delta)^{1/3}})$.

Now, let us see what would happen if we instead employed the above Corollary 13. First, observe that $\sum_{i=1}^d G_i \leq \sqrt{d}\sqrt{\sum_{i=1}^d G_i^2} \leq \sqrt{d}G$. Substituting this expression into Theorem 13 now gives an upper bound on $\|\cdot\|_{1,\delta}$ that is $O(\frac{d^{1/3}G^{2/3}}{(N\delta)^{1/3}})$, which is better than the one we could obtain from Corollary 9 under the same assumptions.

5. From Non-smooth to Smooth Guarantees

Let us now see what our results imply for smooth objectives. The following two propositions show that for smooth F, a (δ, ϵ)-stationary point is automatically a $(0, \epsilon')$-stationary point for some appropriate ϵ'. The proofs are in Appendix E.

Proposition 14. *Suppose that F is H-smooth (that is, ∇F is H-Lipschitz) and x also satisfies $\|\nabla F(\mathbf{x})\|_\delta \leq \epsilon$. Then, $\|\nabla F(\mathbf{x})\| \leq \epsilon + H\delta$.*

Proposition 15. *Suppose that F is J-second-order-smooth (that is, $\|\nabla^2 F(\mathbf{x}) - \nabla^2 F(\mathbf{y})\|_{op} \leq J\|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}$ and $\mathbf{y}$). Suppose also that $\mathbf{x}$ satisfies $\|\nabla F(\mathbf{x})\|_\delta \leq \epsilon$. Then, $\|\nabla F(\mathbf{x})\| \leq \epsilon + \frac{J}{2}\delta^2$.*

Now, recall that Corollary 9 shows that we can find a (δ, ϵ) stationary point in $O(\epsilon^{-3}\delta^{-1})$ iteration. Thus, Proposition 14 implies that by setting $\delta = \epsilon/H$, we can find a $(0, \epsilon)$-stationary point of an H-smooth objective F in $O(\epsilon^{-4})$ iterations, which matches the (optimal) guarantee of standard SGD (Ghadimi & Lan, 2013; Arjevani et al., 2019). Further, Proposition 15 shows that by setting $\delta = \sqrt{\epsilon/J}$, we can find a $(0, \epsilon)$-stationary point of a J-second order smooth objective in $O(\epsilon^{-3.5})$ iterations. This matches the performance of more refined SGD variants and is also known to be tight (Fang et al., 2019; Cutkosky & Mehta, 2020; Arjevani et al., 2020). In summary: **the online-to-non-convex conversion also recovers the optimal results for smooth stochastic losses.**

6. Deterministic and Smooth Case

We will now consider the case of a non-stochastic oracle (that is, $\textsc{Grad}(\mathbf{x}, \mathbf{z}) = \nabla F(\mathbf{x})$ for all $\mathbf{z}$, $\mathbf{x}$) and F is H-smooth (i.e. ∇F is H-Lipschitz). We will show that *optimistic* online algorithms (Rakhlin & Sridharan, 2013; Hazan & Kale, 2010) achieve rates matching the optimal deterministic results. In particular, we consider online algorithms that ensure static regret:

$$R_T(\mathbf{u}) \leq O\left(D\sqrt{\sum_{t=1}^T \|\mathbf{h}_t - \mathbf{g}_t\|^2}\right), \tag{2}$$

for some "hint" vectors $\mathbf{h}_t$. In Appendix B, we provide an explicit construction of such an algorithm for completeness. The standard setting for the hints is $\mathbf{h}_t = \mathbf{g}_{t-1}$. As explained in Section 2.2, to obtain a K-shifting regret it will be enough to reset the algorithm every T iterations.

Theorem 16. *Suppose we have a budget of N gradient evaluations. and that we have an online algorithm $\mathcal{A}_{static}$ that guarantees $\|\mathbf{\Delta}_n\| \leq D$ for all n and ensures the optimistic regret bound $R_T(\mathbf{u}) \leq CD\sqrt{\sum_{t=1}^T \|\mathbf{g}_t - \mathbf{g}_{t-1}\|^2}$ for some constant C, and we define $\mathbf{g}_0 = \mathbf{0}$. In Algorithm 1, set $\mathcal{A}$ to be $\mathcal{A}_{static}$ that is reset every T rounds. Let $\delta > 0$ be an arbitrary number. Set $D = \delta/T$, $T = \min(\lceil\frac{(C\delta^2\sqrt{H}N)^{2/5}}{(F(\mathbf{x}_0)-F^\star)^{2/5}}\rceil, \frac{N}{2})$, and $K = \lfloor\frac{N}{T}\rfloor$. Finally, suppose that F is H-smooth and that the gradient oracle is deterministic (that is, $\mathbf{g}_n = \nabla F(\mathbf{w}_n)$). Then we have:*

$$\left[\frac{1}{K}\sum_{t=1}^K \|\nabla F(\overline{\mathbf{w}}^k)\|_\delta\right] \leq \frac{2CG_1}{N} + \frac{2(F(\mathbf{x}_0) - F^\star)}{\delta N} + \max\left(6\frac{(CH)^{2/5}(F(\mathbf{x}_0) - F^\star)^{3/5}}{\delta^{1/5}N^{3/5}}, \frac{17C\delta\sqrt{H}}{N^{3/2}}\right).$$

Note that the expectation here encompasses only the randomness in the choice of s_t^k, because the gradient oracle is assumed to be deterministic. Theorem 16 finds a (δ, ϵ) stationary point in $O(\epsilon^{-5/3}\delta^{-1/3})$ iteratations. Thus, by setting $\delta = \epsilon/H$, Proposition 14 shows we can find a $(0, \epsilon)$ stationary point in $O(\epsilon^{-2})$ iterations, which matches the standard optimal rate (Carmon et al., 2021).

Proof. First, observe that for all k, t, $\|\overline{\mathbf{w}}^k - \mathbf{w}_t^k\| \leq \delta$. This holds for precisely the same reason that it holds in Corollary 9.

Next, observe that for $k = 1$ we have

$$\begin{aligned}
R_T(\mathbf{u}^k) &\leq CD\sqrt{\sum_{t=1}^T \|\mathbf{g}_t^k - \mathbf{g}_{t-1}^k\|^2} \\
&\leq CD\sqrt{G_1^2 + \sum_{t=2}^T \|\nabla F(\mathbf{w}_t^k) - \nabla F(\mathbf{w}_{t-1}^k)\|^2} \\
&\leq CD\sqrt{G_1^2 + \sum_{t=2}^T H^2\|\mathbf{w}_t^k - \mathbf{w}_{t-1}^k\|^2} \\
&\leq CD\sqrt{G_1^2 + 4H^2TD^2} \leq CDG_1 + 2CD^2H\sqrt{T}\,.
\end{aligned}$$

Similarly, for $k > 1$, we observe that

$$\begin{aligned}
\sum_{t=1}^T \|\mathbf{g}_t^k - \mathbf{g}_{t-1}^k\|^2 &\leq \sum_{t=2}^T \|\nabla F(\mathbf{w}_t^k) - \nabla F(\mathbf{w}_{t-1}^k)\|^2 \\
&\quad + \|\nabla F(\mathbf{w}_1^k) - \nabla F(\mathbf{w}_T^{k-1})\|^2 \\
&\leq H^2(\|\mathbf{w}_1^k - \mathbf{w}_T^{k-1}\|^2 + \sum_{t=2}^T \|\mathbf{w}_t^k - \mathbf{w}_{t-1}^k\|^2) \\
&\leq 4TH^2D^2\,.
\end{aligned}$$

Thus, we have

$$\begin{aligned}
R_T(\mathbf{u}^k) &\leq CD\sqrt{\sum_{t=1}^T \|\mathbf{g}_t^k - \mathbf{g}_{t-1}^k\|^2} \\
&\leq CD\sqrt{4H^2TD^2} \\
&\leq 2CD^2H\sqrt{T}\,.
\end{aligned}$$

Now, applying Theorem 8 in concert with the above bounds on $R_T(\mathbf{u}^k)$, we have

$$\begin{aligned}
&\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^K \left\|\frac{1}{T}\sum_{t=1}^T \nabla F(\mathbf{w}_t^k)\right\|\right] \\
&\leq \frac{2(F(\mathbf{x}_0) - F^\star)}{DN} + \frac{2CDG_1 + 4CKD^2\sqrt{HT}}{DN} \\
&= \frac{2T(F(\mathbf{x}_0) - F^\star)}{\delta N} + \frac{2CG_1}{N} + \frac{4C\delta\sqrt{H}}{T^{3/2}} \\
&\leq \max\left(6\frac{(CH)^{2/5}(F(\mathbf{x}_0) - F^\star)^{3/5}}{\delta^{1/5}N^{3/5}}, \frac{17C\delta\sqrt{H}}{N^{3/2}}\right) \\
&\quad + \frac{2CG_1}{N} + \frac{2(F(\mathbf{x}_0) - F^\star)}{\delta N}\,.
\end{aligned}$$

Recalling that $\|\mathbf{w}_t^k - \overline{\mathbf{w}}^k\| \leq \delta$, the conclusion follows. □

Algorithm 2 Optimistic Mirror Descent with Careful Hints

Input: Learning rate η, number Q (Q will be $O(\log N)$), function F, horizon T, radius D
Receive initial iterate $\mathbf{x}_0$
Set $\mathbf{\Delta}_1' = \mathbf{0}$
for $t = 1 \dots T$ **do**
  Set $\mathbf{h}_t^0 = \nabla F(\mathbf{x}_{t-1})$
  for $i = 1 \dots Q$ **do**
    Set $\mathbf{h}_t^i = \nabla F\left(\mathbf{x}_{t-1} + \frac{1}{2}\Pi_{\|\mathbf{\Delta}\|\leq D}\left[\mathbf{\Delta}_t' - \eta\mathbf{h}_t^{i-1}\right]\right)$
  end for
  Set $\mathbf{h}_t = \mathbf{h}_t^Q$
  Output $\mathbf{\Delta}_t = \Pi_{\|\mathbf{\Delta}\|\leq D}[\mathbf{\Delta}_t' - \eta\mathbf{h}_t]$
  Receive tth gradient $\mathbf{g}_t$
  Set $\mathbf{\Delta}_{t+1}' = \Pi_{\|\mathbf{\Delta}\|\leq D}[\mathbf{\Delta}_t' - \eta\mathbf{g}_t]$
end for

6.1. Better Results with Second-Order Smoothness

When F is J-second-order smooth (i.e., $\nabla^2 F$ is J-Lipschitz) we can do even better. First, observe that by Theorem 16, if F is J-second-order-smooth, then by Proposition 15, the $O(\epsilon^{-5/3}\delta^{-1/3})$ iteration complexity of Theorem 16 implies an $O(\epsilon^{-11/6})$ iteration complexity for finding $(0, \epsilon)$ stationary points by setting $\delta = \sqrt{\epsilon/J}$. This already improves upon the $O(\epsilon^{-2})$ result for smooth losses, but we can improve still further. The key idea is to generate more informative hints $\mathbf{h}_t$. If we can make $\mathbf{h}_t \approx \mathbf{g}_t$, then by (2), we can achieve smaller regret and so a better guarantee.

To do so, we abandon randomization: instead of choosing s_n randomly, we just set $s_n = 1/2$. This setting still allows $F(\mathbf{x}_n) \approx F(\mathbf{x}_{n-1}) + \langle \mathbf{g}_n, \mathbf{\Delta}_n\rangle$ with very little error when F is second-order-smooth. By inspecting the optimistic mirror descent update formula, we can identify an $\mathbf{h}_t$ with $\|\mathbf{h}_t - \mathbf{g}_t\| \leq O(1/\sqrt{N})$ using $O(\log(N))$ gradient queries. This more advanced online learning algorithm is presented in Algorithm 2 (full analysis in Appendix C).

Overall, Algorithm 2's update has an "implicit" flavor:

$$\begin{aligned}
\mathbf{\Delta}_n &= \Pi_{\|\mathbf{\Delta}\|\leq D}\left[\mathbf{\Delta}_{n-1} - \frac{\mathbf{g}_n}{2H}\right], \\
\mathbf{g}_n &= \nabla F(\mathbf{x}_{n-1} + \mathbf{\Delta}_n/2)\,.
\end{aligned}$$

With this refined online algorithm, we can show the following convergence guarantee, whose proof is in Appendix D.

Theorem 17. *In Algorithm 1, assume that $\mathbf{g}_n = \nabla F(\mathbf{w}_n)$, and set $s_n = \frac{1}{2}$. Use Algorithm 2 restarted every T rounds as $\mathcal{A}$. Let $\delta > 0$ an arbitrary number. Set $T = \min(\lceil\frac{(\delta^2(H+J\delta)N)^{1/3}}{(F(\mathbf{x}_0)-F^\star)^{1/3}}\rceil, N/2)$ and $K = \lfloor\frac{N}{T}\rfloor$. In Algorithm 2, set $\eta = 1/2H$, $D = \delta/T$, and $Q = \lceil\log_2(\sqrt{NG/HD})\rceil$. Finally, suppose that F is J-second-order-smooth. Then, the following facts hold:*

1. *For all k, t, $\|\overline{\mathbf{w}}^k - \mathbf{w}_t^k\| \leq \delta$.*
2. *We have the inequality*
$$\frac{1}{K}\sum_{k=1}^{K}\left\|\frac{1}{T}\sum_{t=1}^{T}\nabla F(\mathbf{w}_t^k)\right\| \leq \frac{4G}{N} + \frac{2(F(\mathbf{x}_0)-F^\star)}{N\delta} + 3\frac{(H+J\delta)^{1/3}(F(\mathbf{x}_0)-F^\star)^{2/3}}{\delta^{1/3}N^{2/3}} + 10\frac{\delta(H+J\delta)}{N^2}.$$
3. *With $\delta = \frac{H^{1/7}(F(\mathbf{x}_0)-F(\mathbf{x}_N))^{2/7}}{J^{3/7}N^{2/7}}$, we have*
$$\frac{1}{K}\sum_{t=1}^{K}\|\nabla F(\overline{\mathbf{w}}^k)\| \leq O\left(\frac{J^{1/7}H^{2/7}(F(\mathbf{x}_0)-F^\star)^{4/7}}{N^{4/7}}\right).$$
Moreover, the total number of gradient queries consumed is $NQ = O(N\log(N))$

This result finds a (δ, ϵ) stationary point in $\tilde{O}(\epsilon^{-3/2}\delta^{-1/2})$ iterations. Via Proposition 15, this translates to $\tilde{O}(\epsilon^{-7/4})$ iterations for finding a $(0, \epsilon)$ stationary point, matching the best known rate (up to a logarithmic factor) (Carmon et al., 2017). Note that this may not be optimal: the best lower bound is $\Omega(\epsilon^{-12/7})$ (Carmon et al., 2021). Intriguingly, our technique seems distinct from previous work, which usually relies on acceleration and detecting or exploiting negative curvature (Carmon et al., 2017; Agarwal et al., 2016; Carmon et al., 2018; Li & Lin, 2022).

7. Lower Bounds

In this section, we show that our $O(\epsilon^{-3}\delta^{-1})$ complexity achieved in Corollary 9 is tight. We do this by a simple extension of the lower bound for stochastic *smooth* non-convex optimization of Arjevani et al. (2019). We provide an informal statement and proof-sketch below. The formal result (Theorem 28) and proof is provided in Appendix F.

Theorem 18 (informal)**.** *There is a universal constant C such that for any δ, ϵ, γ and $G \geq C\frac{\sqrt{\epsilon\gamma}}{\sqrt{\delta}}$, for any first-order algorithm $\mathcal{A}$, there is a G-Lipschitz, C^∞ function $F : \mathbb{R}^d \to \mathbb{R}$ for some d with $F(0) - \inf_{\mathbf{x}} F(\mathbf{x}) \leq \gamma$ and a stochastic first-order gradient oracle for F whose outputs $\mathbf{g}$ satisfy $\mathbb{E}[\|\mathbf{g}\|^2] \leq G^2$ such that such that $\mathcal{A}$ requires $\Omega(G^2\gamma/\delta\epsilon^3)$ stochastic oracle queries to identify a point $\mathbf{x}$ with $\mathbb{E}[\|\nabla F(\mathbf{x})\|_\delta] \leq \epsilon$.*

Proof sketch. The construction of Arjevani et al. (2019) provides, for any σ a function F and stochastic oracle whose outputs have norm at most σ such that F is H-smooth, $O(\sqrt{H\gamma})$-Lipschitz and $\mathcal{A}$ requires $\Omega(\sigma^2 H\gamma/\epsilon^4)$ oracle queries to find a point $\mathbf{x}$ with $\|\nabla F(\mathbf{x})\| \leq 2\epsilon$. By setting $H = \frac{\epsilon}{\delta}$ and $\sigma = G/\sqrt{2}$, this becomes an $\sqrt{\epsilon\gamma/\delta}$-Lipschitz function, and so is at most $G/\sqrt{2}$-Lipschitz. Thus, the second moment of the gradient oracle is at most $G^2/2 + G^2/2 = G^2$. Further, the algorithm requires $\Omega(G^2\gamma/\delta\epsilon^3)$ queries to find a point $\mathbf{x}$ with $\|\nabla F(\mathbf{x})\| \leq 2\epsilon$. Now, if $\|\nabla F(\mathbf{x})\|_\delta \leq \epsilon$, then since F is $H = \frac{\epsilon}{\delta}$-smooth, by Proposition 14, $\|\nabla F(\mathbf{x})\| \leq \epsilon + \delta H = 2\epsilon$. Thus, we see that we need $\Omega(G^2\gamma/\delta\epsilon^3)$ queries to find a point with $\|\nabla F(\mathbf{x})\|_\delta \leq \epsilon$ as desired. □

8. Conclusion

We have presented a new online-to-non-convex conversion technique that applies online learning algorithms to non-convex and non-smooth stochastic optimization. When used with online gradient descent, this achieves the optimal $\epsilon^{-3}\delta^{-1}$ complexity for finding (δ, ϵ) stationary points.

These results suggest new directions for work in online learning. Much past work is motivated by the online-to-batch conversion relating *static* regret to *convex* optimization. We employ *switching* regret for *non-convex* optimization. More refined analysis may be possible via generalizations such as strongly adaptive or dynamic regret (Daniely et al., 2015; Jun et al., 2017; Zhang et al., 2018; Jacobsen & Cutkosky, 2022; Cutkosky, 2020; Lu et al., 2022; Luo et al., 2022; Zhang et al., 2021; Baby & Wang, 2022; Zhang et al., 2022). Moreover, our analysis assumes perfect tuning of constants (e.g., D, T, K) for simplicity. In practice, we would prefer to adapt to unknown parameters, motivating new applications and problems for *adaptive* online learning, which is already an area of active current investigation (see, e.g., Orabona & Pál, 2015; Hoeven et al., 2018; Cutkosky & Orabona, 2018; Cutkosky, 2019; Mhammedi & Koolen, 2020; Chen et al., 2021; Sachs et al., 2022; Zhang & Cutkosky, 2022; Wang et al., 2022). We hope that this expertise can be applied in the non-convex setting as well.

Finally, our results leave an important question unanswered: the current best-known algorithm for *deterministic* non-smooth optimization still requires $O(\epsilon^{-3}\delta^{-1})$ iterations to find a (δ, ϵ)-stationary point (Zhang et al., 2020b). We achieve this same result even in the stochastic case. Thus it is natural to wonder if the deterministic rate is tight. For example, is the $O(\epsilon^{-3/2}\delta^{-1/2})$ complexity we achieve in the smooth setting also achievable in the non-smooth setting? Intriguingly, prior work (Kornowski & Shamir, 2022a; Jordan et al., 2022) shows that randomization is necessary, even if the gradient oracle itself is deterministic.

Acknowledgements

A.C. is supported by NSF grant CCF-2211718, a Google gift and an Amazon Research Award. F.O. is supported by NSF grants 2022446 "Foundations of Data Science Institute" and 2046096 "CAREER: Parameter-free Optimization Algorithms for Machine Learning".

References

Agarwal, N., Allen-Zhu, Z., Bullins, B., Hazan, E., and Ma, T. Finding approximate local minima for nonconvex optimization in linear time. *arXiv preprint arXiv:1611.01146*, 2016.

Allen-Zhu, Z. Natasha 2: Faster non-convex optimization than SGD. In *Advances in neural information processing systems*, pp. 2675–2686, 2018.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Srebro, N., and Woodworth, B. Lower bounds for non-convex stochastic optimization. *arXiv preprint arXiv:1912.02365*, 2019.

Arjevani, Y., Carmon, Y., Duchi, J. C., Foster, D. J., Sekhari, A., and Sridharan, K. Second-order information in non-convex stochastic optimization: Power and limitations. In *Conference on Learning Theory*, pp. 242–299, 2020.

Baby, D. and Wang, Y.-X. Optimal dynamic regret in proper online learning with strongly convex losses and beyond. In *International Conference on Artificial Intelligence and Statistics*, pp. 1805–1845. PMLR, 2022.

Bertsekas, D. P. Stochastic optimization problems with non-differentiable cost functionals. *Journal of Optimization Theory and Applications*, 12(2):218–231, 1973.

Bianchi, P., Hachem, W., and Schechtman, S. Convergence of constant step stochastic gradient descent for non-smooth non-convex functions. *Set-Valued and Variational Analysis*, 30(3):1117–1147, 2022.

Bubeck, S. Convex optimization: Algorithms and complexity. *Foundations and Trends® in Machine Learning*, 8 (3-4):231–357, 2015.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. "convex until proven guilty": Dimension-free acceleration of gradient descent on non-convex functions. In *International Conference on Machine Learning*, pp. 654–663. PMLR, 2017.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Accelerated methods for nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1751–1772, 2018.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Lower bounds for finding stationary points I. *Mathematical Programming*, pp. 1–50, 2019.

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. Lower bounds for finding stationary points II: first-order methods. *Mathematical Programming*, 185(1-2), 2021.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge University Press, 2006.

Cesa-Bianchi, N., Conconi, A., and Gentile, C. On the generalization ability of on-line learning algorithms. *Information Theory, IEEE Transactions on*, 50(9):2050–2057, 2004.

Chen, L., Luo, H., and Wei, C.-Y. Impossible tuning made possible: A new expert algorithm and its applications. In *Conference on Learning Theory*, pp. 1216–1259. PMLR, 2021.

Clarke, F. H. *Optimization and nonsmooth analysis*. SIAM, 1990.

Cutkosky, A. Combining online learning guarantees. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pp. 895–913, 2019.

Cutkosky, A. Parameter-free, dynamic, and strongly-adaptive online learning. In *International Conference on Machine Learning*, volume 2, 2020.

Cutkosky, A. and Mehta, H. Momentum improves normalized SGD. In *International Conference on Machine Learning*, 2020.

Cutkosky, A. and Orabona, F. Black-box reductions for parameter-free online learning in Banach spaces. In *Conference On Learning Theory*, pp. 1493–1529, 2018.

Cutkosky, A. and Orabona, F. Momentum-based variance reduction in non-convex SGD. In *Advances in Neural Information Processing Systems*, pp. 15210–15219, 2019.

Daniely, A., Gonen, A., and Shalev-Shwartz, S. Strongly adaptive online learning. In *International Conference on Machine Learning*, pp. 1405–1411. PMLR, 2015.

Davis, D., Drusvyatskiy, D., Lee, Y. T., Padmanabhan, S., and Ye, G. A gradient sampling method with complexity guarantees for lipschitz functions in high and low dimensions. *arXiv preprint arXiv:2112.06969*, 2021.

Duchi, J., Hazan, E., and Singer, Y. Adaptive subgradient methods for online learning and stochastic optimization. In *Conference on Learning Theory (COLT)*, pp. 257–269, 2010.

Duchi, J. C., Bartlett, P. L., and Wainwright, M. J. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. SPIDER: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pp. 689–699, 2018.

Fang, C., Lin, Z., and Zhang, T. Sharp analysis for nonconvex sgd escaping from saddle points. In *Conference on Learning Theory*, pp. 1192–1234, 2019.

Faw, M., Tziotis, I., Caramanis, C., Mokhtari, A., Shakkottai, S., and Ward, R. The power of adaptivity in sgd: Self-tuning step sizes with unbounded gradients and affine variance. In *Conference on Learning Theory*, pp. 313–355. PMLR, 2022.

Flaxman, A. D., Kalai, A. T., and McMahan, H. B. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pp. 385–394, 2005.

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Ghai, U., Lu, Z., and Hazan, E. Non-convex online learning via algorithmic equivalence. *arXiv preprint arXiv:2205.15235*, 2022.

Goyal, P., Dollár, P., Girshick, R., Noordhuis, P., Wesolowski, L., Kyrola, A., Tulloch, A., Jia, Y., and He, K. Accurate, large minibatch SGD: Training ImageNet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.

Hazan, E. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

Hazan, E. and Kale, S. Extracting certainty from uncertainty: Regret bounded by variation in costs. *Machine learning*, 80(2-3):165–188, 2010.

Hazan, E., Singh, K., and Zhang, C. Efficient regret minimization in non-convex games. In *International Conference on Machine Learning*, pp. 1433–1441. PMLR, 2017.

Herbster, M. and Warmuth, M. K. Tracking the best regressor. In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 24–31, 1998.

Hoeven, D., Erven, T., and Kotłowski, W. The many faces of exponential weights in online learning. In *Conference On Learning Theory*, pp. 2067–2092. PMLR, 2018.

Jacobsen, A. and Cutkosky, A. Parameter-free mirror descent. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pp. 4160–4211. PMLR, 2022.

Jordan, M. I., Lin, T., and Zampetakis, M. On the complexity of deterministic nonsmooth and nonconvex optimization. *arXiv preprint arXiv:2209.12463*, 2022.

Jun, K.-S., Orabona, F., Wright, S., and Willett, R. Improved strongly adaptive online learning using coin betting. In *Artificial Intelligence and Statistics*, pp. 943–951. PMLR, 2017.

Karimireddy, S. P., Jaggi, M., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., and Suresh, A. T. Mime: Mimicking centralized stochastic algorithms in federated learning. *arXiv preprint arXiv:2008.03606*, 2020.

Kingma, D. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kornowski, G. and Shamir, O. On the complexity of finding small subgradients in nonsmooth optimization. In *OPT 2022: Optimization for Machine Learning (NeurIPS 2022 Workshop)*, 2022a.

Kornowski, G. and Shamir, O. Oracle complexity in nonsmooth nonconvex optimization. *Journal of Machine Learning Research*, 23(314):1–44, 2022b.

Levy, K., Kavis, A., and Cevher, V. Storm+: Fully adaptive SGD with recursive momentum for nonconvex optimization. *Advances in Neural Information Processing Systems*, 34:20571–20582, 2021.

Li, H. and Lin, Z. Restarted nonconvex accelerated gradient descent: No more polylogarithmic factor in the $o(\epsilon^{-7/4})$ complexity. In *International Conference on Machine Learning*. PMLR, 2022.

Li, X. and Orabona, F. On the convergence of stochastic gradient descent with adaptive stepsizes. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 983–992. PMLR, 2019.

Li, X., Zhuang, Z., and Orabona, F. A second look at exponential and cosine step sizes: Simplicity, adaptivity, and performance. In *International Conference on Machine Learning*, pp. 6553–6564. PMLR, 2021.

Lin, T., Zheng, Z., and Jordan, M. Gradient-free methods for deterministic and stochastic nonsmooth nonconvex optimization. *Advances in Neural Information Processing Systems*, 35:26160–26175, 2022.

Liu, Z., Nguyen, T. D., Nguyen, T. H., Ene, A., and Nguyen, H. L. META-STORM: Generalized fully-adaptive variance reduced SGD for unbounded functions. *arXiv preprint arXiv:2209.14853*, 2022.

Loshchilov, I. and Hutter, F. SGDR: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

Lu, Z., Xia, W., Arora, S., and Hazan, E. Adaptive gradient methods with local guarantees. *arXiv preprint arXiv:2203.01400*, 2022.

Luo, H., Zhang, M., Zhao, P., and Zhou, Z.-H. Corralling a larger band of bandits: A case study on switching regret for linear bandits. In *Conference on Learning Theory*, 2022.

McMahan, H. B. and Streeter, M. Adaptive bound optimization for online convex optimization. In *Proceedings of the 23rd Annual Conference on Learning Theory (COLT)*, pp. 244–256, 2010.

Mhammedi, Z. and Koolen, W. M. Lipschitz and comparator-norm adaptivity in online learning. *Conference on Learning Theory*, pp. 2858–2887, 2020.

Orabona, F. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

Orabona, F. and Pál, D. Scale-free algorithms for online linear optimization. In Chaudhuri, K., Gentile, C., and Zilles, S. (eds.), *Algorithmic Learning Theory*, pp. 287–301. Springer International Publishing, 2015.

Patel, V. and Berahas, A. S. Gradient descent in the absence of global lipschitz continuity of the gradients: Convergence, divergence and limitations of its continuous approximation. *arXiv preprint arXiv:2210.02418*, 2022.

Rakhlin, A. and Sridharan, K. Online learning with predictable sequences. In *Conference on Learning Theory (COLT)*, pp. 993–1019, 2013.

Sachs, S., Hadiji, H., van Erven, T., and Guzmán, C. Between stochastic and adversarial online convex optimization: Improved regret bounds via smoothness. *arXiv preprint arXiv:2202.07554*, 2022.

Stein, E. M. and Shakarchi, R. *Real analysis: measure theory, integration, and Hilbert spaces*. Princeton University Press, 2009.

Tian, L. and So, A. M.-C. No dimension-free deterministic algorithm computes approximate stationarities of lipschitzians. *arXiv preprint arXiv:2210.06907*, 2022.

Tian, L., Zhou, K., and So, A. M.-C. On the finite-time complexity and practical computation of approximate stationarity concepts of lipschitz functions. In *International Conference on Machine Learning*, pp. 21360–21379. PMLR, 2022.

Tripuraneni, N., Stern, M., Jin, C., Regier, J., and Jordan, M. I. Stochastic cubic regularization for fast nonconvex optimization. In *Advances in neural information processing systems*, pp. 2899–2908, 2018.

Wang, G., Hu, Z., Muthukumar, V., and Abernethy, J. Adaptive oracle-efficient online learning. *arXiv preprint arXiv:2210.09385*, 2022.

You, Y., Li, J., Reddi, S., Hseu, J., Kumar, S., Bhojanapalli, S., Song, X., Demmel, J., Keutzer, K., and Hsieh, C.-J. Large batch optimization for deep learning: Training BERT in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.

Zhang, J. and Cutkosky, A. Parameter-free regret in high probability with heavy tails. In *Advances in Neural Information Processing Systems*, 2022.

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S., Kumar, S., and Sra, S. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020a.

Zhang, J., Lin, H., Jegelka, S., Sra, S., and Jadbabaie, A. Complexity of finding stationary points of nonconvex nonsmooth functions. In *International Conference on Machine Learning*, 2020b.

Zhang, L., Lu, S., and Zhou, Z.-H. Adaptive online learning in dynamic environments. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 1330–1340, 2018.

Zhang, L., Wang, G., Tu, W.-W., Jiang, W., and Zhou, Z.-H. Dual adaptivity: A universal algorithm for minimizing the adaptive regret of convex functions. *Advances in Neural Information Processing Systems*, 34:24968–24980, 2021.

Zhang, Z., Cutkosky, A., and Paschalidis, I. Adversarial tracking control via strongly adaptive online learning with memory. In *International Conference on Artificial Intelligence and Statistics*, pp. 8458–8492. PMLR, 2022.

Zhou, D., Xu, P., and Gu, Q. Stochastic nested variance reduction for nonconvex optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 3925–3936, 2018.

Zhuang, Z., Cutkosky, A., and Orabona, F. Surrogate losses for online learning of stepsizes in stochastic non-convex optimization. In *International Conference on Machine Learning*, pp. 7664–7672, 2019.

Zinkevich, M. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pp. 928–936, 2003.

A. Proof of Proposition 2

First, we state a technical lemma that will be used to prove Proposition 2.

Lemma 19. *Let $F : \mathbb{R}^d \to \mathbb{R}$ be locally Lipschitz. Then, F is differentiable almost everywhere, is Lipschitz on all compact sets, and for all $\mathbf{v} \in \mathbb{R}^d$, $\mathbf{x} \mapsto \langle \nabla F(\mathbf{x}), \mathbf{v}\rangle$ is integrable on all compact sets. Finally, for any compact measurable set $D \subset \mathbb{R}^d$, the vector $\mathbf{w} = \int_D \nabla F(\mathbf{x}) \,\mathrm{d}\mathbf{x}$ is well-defined and the operator $\rho(\mathbf{v}) = \int_D \langle \nabla F(\mathbf{x}), \mathbf{v}\rangle \,\mathrm{d}\mathbf{x}$ is linear and equal to $\langle \mathbf{w}, \mathbf{v}\rangle$.*

Proof. First, observe that since F is locally Lipschitz, for every point $\mathbf{x} \in \mathbb{R}^d$ with rational coordinates there is a neighborhood $U_{\mathbf{x}}$ of $\mathbf{x}$ on which F is Lipschitz. Thus, by Rademacher's theorem, F is differentiable almost everywhere in $U_{\mathbf{x}}$. Since the set of points with rational coordiantes is dense in $\mathbb{R}^d$, $\mathbb{R}^d$ is equal to the countable union $\bigcup U_{\mathbf{x}}$. Thus, since the set of points of non-differentiability of F in $U_{\mathbf{x}}$ is measure zero, the total set of points of non-differentability is a countable union of sets of measure zero and so must be measure zero. Thus F is differentiable almost everywhere. This implies that F is differentiable at $\mathbf{x} + p\mathbf{u}$ with probability 1.

Next, observe that for any compact set $S \subset \mathbb{R}^d$, for every point $\mathbf{x} \in S$ with rational coordinates, F is Lipschitz on some neighborhood $U_{\mathbf{x}}$ containing $\mathbf{x}$ with Lipschitz constant $G_{\mathbf{x}}$. Since S is compact, there is a finite set $\mathbf{x}_1, \dots, \mathbf{x}_K$ such that $S = \bigcup U_{\mathbf{x}_i}$. Therefore, F is $\max_i G_{\mathbf{x}_i}$-Lipschitz on S and so F is Lipschitz on every compact set.

Now, for almost all $\mathbf{x}$, for all $\mathbf{v}$ we have that the limit $\lim_{\delta\to 0} \frac{F(\mathbf{x}+\delta\mathbf{v})-F(\mathbf{x})}{\delta}$ exists and is equal to $\langle \nabla F(\mathbf{x}), \mathbf{v}\rangle$ by definition of differentiability. Further, on any compact set S we have that $\frac{|F(\mathbf{x}+\delta\mathbf{v})-F(\mathbf{x})|}{\delta} \le L$ for some L for all $\mathbf{x} \in S$. Therefore, by the bounded convergence theorem (see e.g., Stein & Shakarchi (2009, Theorem 1.4)), we have that $\langle \nabla F(\mathbf{x}), \mathbf{v}\rangle$ is integrable on S.

Next, we prove the linearity of the operator ρ. Observe that for any vectors $\mathbf{v}$ and $\mathbf{w}$, and scalar c, by linearity of integration, we have

$$\begin{aligned}\int_D \langle \nabla F(\mathbf{x}), c\mathbf{v} + \mathbf{w}\rangle \,\mathrm{d}\mathbf{x} &= \int_D \langle \nabla F(\mathbf{x}), c\mathbf{v}\rangle + \langle \nabla F(\mathbf{x}), \mathbf{w}\rangle \,\mathrm{d}\mathbf{x} \\ &= \int_D c\langle \nabla F(\mathbf{x}), \mathbf{v}\rangle + \langle \nabla F(\mathbf{x}), \mathbf{w}\rangle \,\mathrm{d}\mathbf{x} \\ &= c\int_D \langle \nabla F(\mathbf{x}), \mathbf{v}\rangle \,\mathrm{d}\mathbf{x} + \int_D \langle \nabla F(\mathbf{x}), \mathbf{w}\rangle \,\mathrm{d}\mathbf{x}\,.\end{aligned}$$

For the remaining statement, given that $\mathbb{R}^d$ is finite dimensional, there must exist $\mathbf{w} \in \mathbb{R}^d$ such that $\rho(\mathbf{v}) = \langle \mathbf{w}, \mathbf{v}\rangle$ for all $\mathbf{v}$. Further, $\mathbf{w}$ is uniquely determined by $\langle \mathbf{w}, \mathbf{e}_i\rangle$ for $i = 1, \dots, d$ where $\mathbf{e}_i$ indicates the ith standard basis vector. Then, since $\nabla F(\mathbf{x})_i = \langle \nabla F(\mathbf{x}), \mathbf{e}_i\rangle$ is integrable on compact sets, we have

$$\langle \mathbf{w}, \mathbf{e}_i\rangle = \int_D \langle \nabla F(\mathbf{x}), \mathbf{e}_i\rangle \,\mathrm{d}\mathbf{x} = \int_D \nabla F(\mathbf{x})_i \,\mathrm{d}\mathbf{x},$$

which is the definition of $\int_D \nabla F(\mathbf{x}) \,\mathrm{d}\mathbf{x}$ when the integral is defined. □

We can now prove the Proposition.

Proof of Proposition 2. Since F is locally Lipschitz, by Proposition 19, F is Lipschitz on compact sets. Therefore, F must be Lipschitz on the line segment connecting $\mathbf{x}$ and $\mathbf{y}$. Thus the function $k(t) = F(\mathbf{x} + t(\mathbf{y} - \mathbf{x}))$ is absolutely continuous on $[0, 1]$. As a result, k' is integrable on $[0, 1]$ and $F(\mathbf{y}) - F(\mathbf{x}) = k(1) - k(0) = \int_0^1 k'(t) \,\mathrm{d}t = \int_0^1 \langle \nabla F(\mathbf{x}+t(\mathbf{y}-\mathbf{x})), \mathbf{y}-\mathbf{x}\rangle \,\mathrm{d}t$ by the Fundamental Theorem of Calculus (see, e.g., Stein & Shakarchi (2009, Theorem 3.11)).

Now, we tackle the case that F is not differentiable everywhere. Notice that the last statement of the Proposition is an immediate consequence of Lipschitzness. So, we focus on showing the remaining parts.

Now, by Lemma 19, we have that F is differentiable almost everywhere. Further, $\mathbf{g}_{\mathbf{x}} = \mathbb{E}_{\mathbf{u}}[\nabla F(\mathbf{x} + p\mathbf{u})]$ exists and satisfies for all $\mathbf{v} \in \mathbb{R}^d$:

$$\langle \mathbf{g}_{\mathbf{x}}, \mathbf{v}\rangle = \mathop{\mathbb{E}}_{\mathbf{u}}[\langle \nabla F(\mathbf{x} + p\mathbf{u}), \mathbf{v}\rangle]\,.$$

Notice also that

$$\mathop{\mathbb{E}}_{\mathbf{z},\mathbf{u}}\left[\hat{\textsc{Grad}}(\mathbf{x},(\mathbf{z},\mathbf{u}))\right] = \mathop{\mathbb{E}}_{\mathbf{u}}\mathop{\mathbb{E}}_{\mathbf{z}}[\textsc{Grad}(\mathbf{x}+p\mathbf{u},\mathbf{z})] = \mathop{\mathbb{E}}_{\mathbf{u}}[\nabla F(\mathbf{x}+p\mathbf{u})] = \mathbf{g}_{\mathbf{x}} \ .$$

So, it remains to show that $\hat{F}$ is differentiable and $\nabla\hat{F}(\mathbf{x}) = \mathbf{g}_{\mathbf{x}}$.

Now, let $\mathbf{x}$ be an arbitrary elements of $\mathbb{R}^d$ and let $\mathbf{v}_1, \mathbf{v}_2, \dots$ be any sequence of vectors such that $\lim_{n\to\infty}\mathbf{v}_n = 0$ and $\|\mathbf{v}_i\| \le p$ for all i. Then, since the ball of radius $2p$ centered at $\mathbf{x}$ is compact, F is L-Lipschitz inside this ball for some L. Then, we have

$$\lim_{n\to\infty}\frac{\hat{F}(\mathbf{x}+\mathbf{v}_n)-\hat{F}(\mathbf{x})-\langle\mathbf{g}_{\mathbf{x}},\mathbf{v}_n\rangle}{\|\mathbf{v}_n\|} = \lim_{n\to\infty}\mathop{\mathbb{E}}_{u}\left[\frac{F(\mathbf{x}+\mathbf{v}_n+p\mathbf{u})-F(\mathbf{x}+p\mathbf{u})-\langle\nabla F(\mathbf{x}+p\mathbf{u}),\mathbf{v}_n\rangle}{\|\mathbf{v}_n\|}\right] \ .$$

Now, observe $\frac{|F(\mathbf{x}+\mathbf{v}_i+p\mathbf{u})-F(\mathbf{x}+p\mathbf{u})|}{\|\mathbf{v}_i\|} \le L$. Further, for all almost all $\mathbf{u}$, F is differentiable at $\mathbf{x}+p\mathbf{u}$ so that $\lim_{n\to\infty}\frac{F(\mathbf{x}+\mathbf{v}_n+p\mathbf{u})-F(\mathbf{x}+p\mathbf{u})-\langle\nabla F(\mathbf{x}+p\mathbf{u}),\mathbf{v}_n\rangle]}{\|\mathbf{v}_n\|} = 0$ for almost all $\mathbf{u}$. Thus, by the bounded convergence theorem, we have

$$\lim_{n\to\infty}\frac{\hat{F}(\mathbf{x}+\mathbf{v}_n)-\hat{F}(\mathbf{x})-\langle\mathbf{g}_{\mathbf{x}},\mathbf{v}_n\rangle}{\|\mathbf{v}_n\|} = 0 \ .$$

which shows that $\mathbf{g}_{\mathbf{x}} = \nabla\hat{F}(\mathbf{x})$.

Finally, observe that since F is Lipschitz on compact sets, $\hat{F}$ must be also, and so by the first part of the proposition, $\hat{F}$ is well-behaved. □

B. Analysis of (Optimistic) Online Gradient Descent

Optimistic Online Gradient Descent (in its simplest form) is described by Algorithm 3. Here we collect the standard analysis of the algorithm for completeness. None of this analysis is new, and more refined versions can be found in a variety of sources (e.g. (Chen et al., 2021)).

Algorithm 3 Optimistic Mirror Descent

Input: Regularizer function ϕ, domain V, time horizon T.
$\hat{\mathbf{w}}_1 = \mathbf{0}$
for $t = 1\dots T$ **do**
  Generate "hint" h_t
  Set $\mathbf{w}_t = \text{argmin}_{\mathbf{x}\in V}\ \langle\mathbf{h}_t,\mathbf{x}\rangle + \frac{1}{2}\|\mathbf{x}-\hat{\mathbf{w}}_t\|^2$
  Output $\mathbf{w}_t$ and receive loss vector $\mathbf{g}_t$
  Set $\hat{\mathbf{w}}_{t+1} = \text{argmin}_{\mathbf{x}\in V}\ \langle\mathbf{g}_t,\mathbf{x}\rangle + \frac{1}{2}\|\mathbf{x}-\hat{\mathbf{w}}_t\|^2$
end for

We will analyze only a simple version of this algorithm, that is when V is an L_2 ball of radius D in some real Hilbert space (such as $\mathbb{R}^d$). Then, Algorithm 3 satisfies the following guarantee.

Proposition 20. *Let $V = \{\mathbf{x} : \|\mathbf{x}\| \le D\} \subset \mathcal{H}$ for some real Hilbert space $\mathcal{H}$. Then, with for all $\mathbf{u} \in V$, Algorithm 3 ensures*

$$\sum_{t=1}^{T}\langle\mathbf{g}_t,\mathbf{w}_t-\mathbf{u}\rangle \le \frac{D^2}{2\eta} + \sum_{t=1}^{T}\frac{\eta}{2}\|\mathbf{g}_t-\mathbf{h}_t\|^2 \ .$$

Proof. Now, by Chen et al. (2021, Lemma 15) instantiated with the squared Euclidean distance as Bregman divergence, we have

$$\langle\mathbf{g}_t,\mathbf{w}_t-\mathbf{u}\rangle \le \langle\mathbf{g}_t-\mathbf{h}_t,\mathbf{w}_t-\hat{\mathbf{w}}_{t+1}\rangle + \frac{1}{2}\|\mathbf{u}-\hat{w}_t\|^2 - \frac{1}{2}\|\mathbf{u}-\hat{\mathbf{w}}_{t+1}\|^2 - \frac{1}{2}\|\hat{\mathbf{w}}_{t+1}-\mathbf{w}_t\|^2 - \frac{1}{2}\|\mathbf{w}_t-\hat{\mathbf{w}}_t\|^2$$

From Young inequality:

$$\leq \frac{\eta\|\mathbf{g}_t - \mathbf{h}_t\|^2}{2} + \frac{\|\mathbf{w}_t - \hat{\mathbf{w}}_{t+1}\|^2}{2\eta} + \frac{1}{2}\|\mathbf{u} - \hat{\mathbf{w}}_t\|^2 - \frac{1}{2}\|\mathbf{u} - \hat{\mathbf{w}}_{t+1}\|^2 - \frac{1}{2}\|\hat{\mathbf{w}}_{t+1} - \mathbf{w}_t\|^2 - \frac{\|\mathbf{w}_t - \hat{\mathbf{w}}_t\|^2}{2\eta}$$
$$\leq \frac{\eta\|\mathbf{g}_t - \mathbf{h}_t\|^2}{2} + \frac{1}{2}\|\mathbf{u} - \hat{\mathbf{w}}_t\|^2 - \frac{1}{2}\|\mathbf{u} - \hat{\mathbf{w}}_{t+1}\|^2 .$$

Summing over t and telescoping, we have

$$\sum_{t=1}^{T}\langle \mathbf{g}_t, \mathbf{w}_t - \mathbf{u}\rangle \leq \frac{1}{2}\|\mathbf{u} - \hat{\mathbf{w}}_1\|^2 - \frac{1}{2}\|\mathbf{u} - \hat{\mathbf{w}}_{T+1}\|^2 + \sum_{t=1}^{T}\frac{\eta\|\mathbf{g}_t - \mathbf{h}_t\|^2}{2}$$
$$\leq \frac{\|\mathbf{u} - \hat{\mathbf{w}}_1\|^2}{2\eta} + \sum_{t=1}^{T}\frac{\eta\|\mathbf{g}_t - \mathbf{h}_t\|^2}{2} \leq \frac{D^2}{2\eta} + \sum_{t=1}^{T}\frac{\eta\|\mathbf{g}_t - \mathbf{h}_t\|^2}{2} . \qquad \square$$

In the case that the hints $\mathbf{h}_t$ are not present, the algorithm becomes online gradient descent (Zinkevich, 2003). In this case, assuming $\mathbb{E}[\|\mathbf{g}_t\|^2] \leq G^2$ and setting $\eta = \frac{D}{G\sqrt{T}}$ we obtain the $\mathbb{E}[R_T(\mathbf{u})] \leq DG\sqrt{T}$ for all $\mathbf{u}$ such that $\|\mathbf{u}\| \leq D$.

C. Algorithm 2 and Regret Guarantee

Theorem 21. *Let F be an H-smooth and G-Lipschitz function. Then, when $Q = \lceil \log_2(\sqrt{NG/HD})\rceil$, Algorithm 2 with $\mathbf{x}_t = \mathbf{x}_{t-1} + \frac{1}{2}\mathbf{\Delta}_t$ and $\mathbf{g}_t = \nabla F(\mathbf{x}_t)$ and $\eta \leq \frac{1}{2H}$ ensures for all $\|\mathbf{u}\| \leq D$*

$$\sum_{t=1}^{T}\langle \mathbf{g}_t, \mathbf{\Delta}_t - \mathbf{u}\rangle \leq \frac{HD^2}{2} + \frac{2GTD}{N} .$$

Furthermore, a total of at most $T\lceil \log_2(\sqrt{NG/HD})\rceil$ gradient evaluations are required.

Proof. The count of gradient evaluations is immediate from inspection of the algorithm, so it remains only to prove the regret bound.

First, we observe that the choices of $\mathbf{\Delta}_t$ specified by Algorithm 2 correspond to the values of $\mathbf{w}_t$ produced by Algorithm 3 when $\psi(\mathbf{w}) = \frac{1}{2\eta}\|\mathbf{w}\|^2$. This can be verified by direct calculation (recalling that $D_\psi(\mathbf{x}, \mathbf{y}) = \frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\eta}$).

Therefore, by Proposition 20, we have

$$\sum_{t=1}^{T}\langle \mathbf{g}_t, \mathbf{\Delta}_t - \mathbf{u}\rangle \leq \frac{D^2}{2\eta} + \sum_{t=1}^{T}\frac{\eta}{2}\|\mathbf{g}_t - \mathbf{h}_t\|^2 . \tag{3}$$

So, our primary task is to show that $\|\mathbf{g}_t - \mathbf{h}_t\|$ is small. To this end, recall that $\mathbf{g}_t = \nabla F(\mathbf{w}_t) = \nabla F(\mathbf{x}_{t-1} + \mathbf{\Delta}_t/2)$.

Now, we define $h_t^{M+1} = \nabla F\left(\mathbf{x}_{t-1} + \frac{1}{2}\Pi_{\|\mathbf{\Delta}\|\leq D}\left[\mathbf{\Delta}_t' - \eta\mathbf{h}_t^M\right]\right)$ (which simply continues the recursive definition of $\mathbf{h}_t^i$ in Algorithm 2 for one more step). Then, we claim that for all $0 \leq i \leq M$, $\|\mathbf{h}_t^{i+1} - \mathbf{h}_t^i\| \leq \frac{1}{2^i}\|\mathbf{h}_t^1 - \mathbf{h}_t^0\|$. We establish the claim by induction on i. First, for $i = 0$ the claim holds by definition. Now suppose $\|\mathbf{h}_t^i - \mathbf{h}_t^{i-1}\| \leq \frac{1}{2^{i-1}}\|\mathbf{h}_t^1 - \mathbf{h}_t^0\|$ for some i. Then, we have

$$\|\mathbf{h}_t^{i+1} - \mathbf{h}_t^i\| \leq \left\|\nabla F\left(\mathbf{x}_{t-1} + \frac{1}{2}\Pi_{\|\mathbf{\Delta}\|\leq D}\left[\mathbf{\Delta}_t' - \eta\mathbf{h}_t^i\right]\right) - \nabla F\left(\mathbf{x}_{t-1} + \frac{1}{2}\Pi_{\|\mathbf{\Delta}\|\leq D}\left[\mathbf{\Delta}_t' - \eta\mathbf{h}_t^{i-1}\right]\right)\right\|$$

Using the H-smoothness of F:

$$\leq \frac{H}{2}\left\|\Pi_{\|\mathbf{\Delta}\|\leq D}\left[\mathbf{\Delta}_t' - \eta\mathbf{h}_t^i\right] - \Pi_{\|\mathbf{\Delta}\|\leq D}\left[\mathbf{\Delta}_t' - \eta\mathbf{h}_t^{i-1}\right]\right\|$$

Using the fact that projection is a contraction:

$$\leq \frac{H\eta}{2} \left\|\mathbf{h}_t^i - \mathbf{h}_t^{i-1}\right\|$$

Using $\eta \leq \frac{1}{H}$:

$$= \frac{1}{2} \left\|\mathbf{h}_t^i - \mathbf{h}_t^{i-1}\right\|$$

From the induction assumption:

$$\leq \frac{1}{2^i} \|\mathbf{h}_t^1 - \mathbf{h}_t^0\| \ .$$

So that the claim holds.

Now, since $\mathbf{h}_t = \mathbf{h}_t^Q$, we have $\mathbf{\Delta}_t = \Pi_{\|\mathbf{\Delta}\| \leq D} \left[\mathbf{\Delta}_t' - \eta \mathbf{h}_t^{i-1}\right]$. Therefore $g_t = \nabla F(\mathbf{x}_t) = \nabla F(\mathbf{x}_{t-1} + \mathbf{\Delta}_t/2) = \mathbf{h}_t^{Q+1}$. Thus,

$$\|\mathbf{g}_t - \mathbf{h}_t^Q\| = \|\mathbf{h}_t^{Q+1} - \mathbf{h}_t^Q\| \leq \frac{1}{2^Q} \|\mathbf{h}_t^1 - \mathbf{h}_t^0\| \leq \frac{2G}{2^Q},$$

where in the last inequality we used the fact that F is G-Lipschitz. So, for $Q = \lceil \log_2(\sqrt{NG/HD}) \rceil$, we have $\|\mathbf{g}_t - \mathbf{h}_t^Q\| \leq \frac{2G\sqrt{HD}}{\sqrt{NG}}$ for all t. The result now follows by substituting into equation (3). □

D. Proof of Theorem 17

Proof of Theorem 17. Once more, the first part of the result is immediate from the fact that $\|\mathbf{\Delta}_n\| \leq D$. So, we proceed to show the second part.

Define $\mathbf{\nabla}_n = \int_0^1 \nabla F(\mathbf{x}_{n-1} + s\mathbf{\Delta}_n) \, \mathrm{d}s$. Then, we have

$$\begin{aligned}
&\|\langle \mathbf{\nabla}_n - \mathbf{g}_n, \mathbf{\Delta}_n \rangle\| \\
&= \left\| \left\langle \int_0^1 \nabla F(\mathbf{x}_{n-1} + s\mathbf{\Delta}_n) - \nabla F\left(\mathbf{x}_{n-1} + \frac{1}{2}\mathbf{\Delta}_n\right) \mathrm{d}s, \ \mathbf{\Delta}_n \right\rangle \right\| \\
&\leq D \left\| \int_0^1 \nabla F(\mathbf{x}_{n-1} + s\mathbf{\Delta}_n) - \nabla F\left(\mathbf{x}_{n-1} + \frac{1}{2}\mathbf{\Delta}_n\right) \mathrm{d}s \right\| \\
&= D \left\| \int_0^1 \left(\nabla F(\mathbf{x}_{n-1} + s\mathbf{\Delta}_n) - \nabla F\left(\mathbf{x}_{n-1} + \frac{1}{2}\mathbf{\Delta}_n\right) - \nabla^2 F\left(\mathbf{x}_{n-1} + \frac{1}{2}\mathbf{\Delta}_n\right) \mathbf{\Delta}_n (s - 1/2) \right) \right. \\
&\qquad \left. + \nabla^2 F(\mathbf{x}_{n-1} + \frac{1}{2}\mathbf{\Delta}_n) \mathbf{\Delta}_n (s - 1/2) \, \mathrm{d}s \right\|
\end{aligned}$$

(observing that $\int_0^1 s - 1/2 \, \mathrm{d}s = 0$)

$$= D \left\| \int_0^1 \nabla F(\mathbf{x}_{n-1} + s\mathbf{\Delta}_n) - \nabla F(\mathbf{x}_{n-1} + \frac{1}{2}\mathbf{\Delta}_n) - \nabla^2 F\left(\mathbf{x}_{n-1} + \frac{1}{2}\mathbf{\Delta}_n\right) \mathbf{\Delta}_n (s - 1/2) \, \mathrm{d}s \right\|$$

(using second-order smoothness)

$$\leq D \int_0^1 \frac{J}{2} \|\mathbf{\Delta}_n\|^2 (s - 1/2)^2 \, \mathrm{d}s \leq \frac{JD^3}{48} \ .$$

In Theorem 7, set $\mathbf{u}_n$ to be equal to $\mathbf{u}^1$ for the first T iterations, $\mathbf{u}^2$ for the second T iterations and so on. In other words, $\mathbf{u}_n = \mathbf{u}^{\lfloor n/T \rfloor + 1}$ for $n = 1, \ldots, M$. So, we have

$$\begin{aligned}
F(\mathbf{x}_M) - F(\mathbf{x}_0) &= R_T(\mathbf{u}^1, \ldots, \mathbf{u}^K) + \sum_{n=1}^M \langle \mathbf{\nabla}_n - \mathbf{g}_n, \mathbf{\Delta}_n \rangle + \sum_{n=1}^M \langle \mathbf{g}_n, \mathbf{u}_n \rangle \\
&\leq R_T(\mathbf{u}^1, \ldots, \mathbf{u}^K) + \frac{NJD^3}{48} + \sum_{n=1}^M \langle \mathbf{g}_n, \mathbf{u}_n \rangle \ .
\end{aligned}$$

Now, set $\mathbf{u}_k = -D\frac{\sum_{t=1}^T \nabla F(\mathbf{w}_t^k)}{\|\sum_{t=1}^T \nabla F(\mathbf{w}_t^k)\|}$. Then, by Theorem 21, we have that $R_T(\mathbf{u}_k) \leq \frac{HD^2}{2} + \frac{2TGD}{N}$. Therefore:

$$F(\mathbf{x}_M) \leq F(\mathbf{x}_0) + \frac{HD^2K}{2} + 2GD + \frac{MJD^3}{48} - DT\sum_{k=1}^K \left\| \frac{1}{T}\sum_{t=1}^T \nabla F(\mathbf{w}_t^k) \right\| .$$

Hence, we obtain

$$\frac{1}{K}\sum_{k=1}^K \left\| \frac{1}{T}\sum_{t=1}^T \nabla F(\mathbf{w}_t^k) \right\| \leq \frac{F(\mathbf{x}_0) - F(\mathbf{x}_M)}{MD} + \frac{HD}{2T} + \frac{2G}{M} + \frac{JD^2}{48} .$$

Note that from the choice of K and T we have $M = KT \geq N - T \geq N/2$. So, using $D = \delta/T$, we have can upper bound the r.h.s. with

$$\frac{2T(F(\mathbf{x}_0) - F^\star)}{N\delta} + \frac{H\delta}{2T^2} + \frac{4G}{N} + \frac{J\delta^2}{2T^2}$$

and with $T = \min\left(\left\lceil \frac{(\delta^2(H+J\delta)N)^{1/3}}{(F(\mathbf{x}_0)-F^\star)^{1/3}} \right\rceil, N/2\right)$:

$$\leq 3\frac{(H+J\delta)^{1/3}(F(\mathbf{x}_0) - F^\star)^{2/3}}{\delta^{1/3}N^{2/3}} + \frac{4G}{N} + \frac{2(F(\mathbf{x}_0) - F^\star)}{N\delta} + 10\frac{\delta(H+J\delta)}{N^2} .$$

Now, the third fact follows by observing that Proposition 15 implies that

$$\frac{1}{K}\sum_{k=1}^K \left\|\nabla F(\overline{\mathbf{w}}^k)\right\| \leq \frac{1}{K}\sum_{k=1}^K \left\| \frac{1}{T}\sum_{t=1}^T \nabla F(\mathbf{w}_t^k) \right\| + \frac{J\delta^2}{2} .$$

Now, substituting the specified value of δ completes the identity. Finally, the count of number of gradient evaluations is a direct calculation. □

E. Proofs for Section 5

Proposition 14. *Suppose that F is H-smooth (that is, ∇F is H-Lipschitz) and x also satisfies $\|\nabla F(\mathbf{x})\|_\delta \leq \epsilon$. Then, $\|\nabla F(\mathbf{x})\| \leq \epsilon + H\delta$.*

Proof. Let $S \subset B(\mathbf{x}, \delta)$ with $\mathbf{x} = \frac{1}{|S|}\sum_{\mathbf{y}\in S}\mathbf{y}$. By H-smoothness, for all $\mathbf{y} \in S$, $\|\nabla F(\mathbf{y}) - \nabla F(\mathbf{x})\| \leq H\|\mathbf{y} - \mathbf{x}\| \leq H\delta$. Therefore, we have

$$\begin{aligned}\left\| \frac{1}{|S|}\sum_{\mathbf{y}\in S}\nabla F(\mathbf{y}) \right\| &= \left\| \nabla F(\mathbf{x}) + \frac{1}{|S|}\sum_{\mathbf{y}\in S}(\nabla F(\mathbf{y}) - \nabla F(\mathbf{x})) \right\| \\ &\geq \|\nabla F(\mathbf{x})\| - H\delta .\end{aligned}$$

Now, since $\|\nabla F(\mathbf{x})\|_\delta \leq \epsilon$, for any $p > 0$, there is a set S such that $\left\|\frac{1}{|S|}\sum_{\mathbf{y}\in S}\nabla F(\mathbf{y})\right\| \leq \epsilon + p$. Thus, $\|\nabla F(\mathbf{x})\| \leq \epsilon + H\delta + p$ for any $p > 0$, which implies $\|\nabla F(\mathbf{x})\| \leq \epsilon + H\delta$. □

Proposition 15. *Suppose that F is J-second-order-smooth (that is, $\|\nabla^2 F(\mathbf{x}) - \nabla^2 F(\mathbf{y})\|_{op} \leq J\|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}$ and $\mathbf{y}$). Suppose also that $\mathbf{x}$ satisfies $\|\nabla F(\mathbf{x})\|_\delta \leq \epsilon$. Then, $\|\nabla F(\mathbf{x})\| \leq \epsilon + \frac{J}{2}\delta^2$.*

Proof. The proof is similar to that of Proposition 14. Let $S \subset B(\mathbf{x}, \delta)$ with $\mathbf{x} = \frac{1}{|S|}\sum_{\mathbf{y}\in S}\mathbf{y}$. By J-second-order-smoothness, for all $\mathbf{y} \in S$, we have

$$\begin{aligned}\|\nabla F(\mathbf{y}) - \nabla F(\mathbf{x}) - \nabla^2 F(\mathbf{x})(\mathbf{y} - \mathbf{x})\| &= \left\| \int_0^1 (\nabla^2 F(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla^2 F(\mathbf{x}))(\mathbf{y} - \mathbf{x})\,\mathrm{d}t \right\| \\ &\leq \int_0^1 tJ\|\mathbf{y} - \mathbf{x}\|^2\,\mathrm{d}t = \frac{J\|\mathbf{y} - \mathbf{x}\|^2}{2} \leq \frac{J\delta^2}{2} .\end{aligned}$$

Further, since $\frac{1}{|S|}\sum_{\mathbf{y}\in S}\mathbf{y}=\mathbf{x}$, we have $\frac{1}{|S|}\sum_{\mathbf{y}\in S}\nabla^2 F(\mathbf{x})(\mathbf{y}-\mathbf{x})=0$. Therefore, we have

$$\begin{aligned}\left\|\frac{1}{|S|}\sum_{\mathbf{y}\in S}\nabla F(\mathbf{y})\right\| &= \left\|\nabla F(\mathbf{x})+\frac{1}{|S|}\sum_{\mathbf{y}\in S}(\nabla F(\mathbf{y})-\nabla F(\mathbf{x})-\nabla^2 F(\mathbf{x})(\mathbf{y}-\mathbf{x}))\right\| \\ &\geq \|\nabla F(\mathbf{x})\|-\frac{J\delta^2}{2}\,.\end{aligned}$$

Now, since $\|\nabla F(x)\|_\delta \leq \epsilon$, for any $p>0$, there is a set S such that $\left\|\frac{1}{|S|}\sum_{\mathbf{y}\in S}\nabla F(\mathbf{y})\right\| \leq \epsilon+p$. Thus, $\|\nabla F(\mathbf{x})\| \leq \epsilon+\frac{J}{2}\delta^2+p$ for any $p>0$, which implies $\|\nabla F(\mathbf{x})\|\leq \epsilon+\frac{J}{2}\delta^2$. □

F. Lower Bounds

Our lower bounds are constructed via a mild alteration to the arguments of Arjevani et al. (2019) for lower bounds on finding $(0,\epsilon)$-stationary points of *smooth* functions with a stochastic gradient oracle. At a high level, we show that since a δ,ϵ-stationary point of an H-smooth loss is also a $(0, H\delta+\epsilon)$-stationary point, a lower bound on the complexity of the latter implies a lower bound on the of complexity of the former. The lower bound of Arjevani et al. (2019) is proved by constructing a distribution over "hard" functions such that no algorithm can quickly find a $(0,\epsilon)$-stationary point of a random selected function. Unfortunately, these "hard" functions are not Lipschitz. Fortunately, they take the form $F(\mathbf{x})+\eta\|\mathbf{x}\|^2$ where F is Lipschitz and smooth so that the "non-Lipschitz" part is solely contained in the quadratic term. We show that one can replace the quadratic term $\|\mathbf{x}\|^2$ with a Lipschitz function that is quadratic for sufficiently small $\mathbf{x}$ but proportional to $\|\mathbf{x}\|$ for larger values. Our proof consists of carefully reproducing the argument of Arjevani et al. (2019) to show that this modification does not cause any problems. We emphasize that almost all of this development can be found with more detail in Arjevani et al. (2019). We merely restate here the minimum results required to verify our modification to their construction.

F.1. Definitions and Results from Arjevani et al. (2019)

A randomized first-order algorithm is a distribution P_S supported on a set S and a sequence of measurable mappings $A_i(s,\mathbf{g}_1,\ldots,\mathbf{g}_{i-1})\to\mathbb{R}^d$ with $s\in S$ and $\mathbf{g}_i\in\mathbb{R}^d$. Given a stochastic gradient oracle $\textsc{Grad}:\mathbb{R}^d\times Z\to\mathbb{R}^d$, a distribution P_Z supported on Z and an i.i.d. sample $(\mathbf{z}_1,\ldots,\mathbf{z}_n)\sim P_Z$, we define the *iterates* of A recursively by:

$$\begin{aligned}\mathbf{x}_1 &= A_1(s)\\ \mathbf{x}_i &= A_i(s,\textsc{Grad}(\mathbf{x}_1,\mathbf{z}_1),\textsc{Grad}(\mathbf{x}_2,\mathbf{z}_2),\ldots,\textsc{Grad}(\mathbf{x}_{i-1},\mathbf{z}_{i-1}))\,.\end{aligned}$$

So, $\mathbf{x}_i$ is a function of s and $\mathbf{z}_1,\ldots,\mathbf{z}_{i-1}$. We define A_{rand} to be the set of such sequences of mappings.

Now, in the notation of Arjevani et al. (2019), we define the "progress function"

$$\text{prog}_c(\mathbf{x})=\max\{i\ :\ |\mathbf{x}_i|\geq c\}\,.$$

Further, a stochastic gradient oracle $\textsc{Grad}$ can be called a probability-p zero-chain if $\text{prog}_0(\textsc{Grad}(\mathbf{x},\mathbf{z}))=\text{prog}_{1/4}(\mathbf{x})+1$ for all $\mathbf{x}$ with probability at least $1-p$, and $\text{prog}_0(\textsc{Grad}(\mathbf{x},\mathbf{z}))\leq\text{prog}_{1/4}(\mathbf{x})+1$ with probability 1.

Next, let $F_T:\mathbb{R}^T\to\mathbb{R}$ be the function defined by Lemma 2 of Arjevani et al. (2019). Restating their Lemma, this function satisfies:

Lemma 22 (Lemma 2 of (Arjevani et al., 2019))**.** *There exists a function $F_T:\mathbb{R}^T\to\mathbb{R}$ satisfies that satisfies:*

1. $F_T(0)=0$ *and* $\inf F_T(\mathbf{x})\geq-\gamma_0 T$ *for* $\gamma_0=12$.
2. $\nabla F_T(\mathbf{x})$ *is* H_0*-Lipschitz, with* $H_0=152$.
3. *For all* $\mathbf{x}$, $\|\nabla F_T(\mathbf{x})\|_\infty\leq G_0$ *with* $G_0=23$
4. *For all* $\mathbf{x}$, $\text{prog}_0(\nabla F_T(\mathbf{x}))\leq\text{prog}_{1/2}(\nabla F_T(\mathbf{x}))+1$.

5. *If* $prog_1(\mathbf{x}) < T$, *then* $\|\nabla F_T(\mathbf{x})\| \geq |\nabla F_T(\mathbf{x})_{prog_1(\mathbf{x})+1}| \geq 1$.

We associate with this function F_T the stochastic gradient oracle $O_T(\mathbf{x}, z) : \mathbb{R}^T \times \{0,1\} \to \mathbb{R}^d$ where z is Bernoulli(p):

$$\textsc{Grad}_T(\mathbf{x}, \mathbf{z})_i = \begin{cases} \nabla F_T(\mathbf{x})_i, & \text{if } i \neq \text{prog}_{1/4}(\mathbf{x}) \\ \frac{\mathbf{z}\nabla F_T(\mathbf{x})_i}{p}, & \text{if } i = \text{prog}_{1/4}(\mathbf{x}) \end{cases}$$

It is clear that $\mathbb{E}_z[O_T(\mathbf{x}, \mathbf{z})] = \nabla F_T(\mathbf{x})$.

This construction is so far identical to that in Arjevani et al. (2019), and so we have by their Lemma 3:

Lemma 23 (Lemma 3 of Arjevani et al. (2019))**.** $\textsc{Grad}_T$ *is a probability-p zero chain, has variance* $\mathbb{E}[\|\textsc{Grad}_T(\mathbf{x}, \mathbf{z}) - \nabla F_T(\mathbf{x})\|^2] \leq G_0^2/p$, *and* $\|\textsc{Grad}_T(\mathbf{x}, z)\| \leq \frac{G_0}{p} + G_0\sqrt{T}$.

Proof. The probability p zero-chain and variance statements are directly from Arjevani et al. (2019). For the bound on $\|\textsc{Grad}_T\|$, observe that $\textsc{Grad}_T(\mathbf{x}, \mathbf{z}) = \nabla F_T(\mathbf{x})$ in all but one coordinate. In that one coordinate, $\textsc{Grad}_T(\mathbf{x}, \mathbf{z})$ is at most $\frac{\|\nabla F_T(\mathbf{x})\|_\infty}{p} = \frac{G_0}{p}$. Thus, the bound follows by triangle inequality. □

Next, for any matrix $U \in \mathbb{R}^{d\times T}$ with orthonormal columns, we define $F_{T,U} : \mathbb{R}^d \to \mathbb{R}$ by:

$$F_{T,U}(\mathbf{x}) = F_T(U^\top \mathbf{x}) \ .$$

The associated stochastic gradient oracle is:

$$\textsc{Grad}_{T,U}(\mathbf{x}, \mathbf{z}) = U\textsc{Grad}_T(U^\top \mathbf{x}, \mathbf{z}) \ .$$

Now, we restate Lemma 5 of Arjevani et al. (2019):

Lemma 24 (Lemma 5 of Arjevani et al. (2019))**.** *Let* $R > 0$ *and suppose* $A \in \mathrm{A}_{rand}$ *is such that* A *produces iterates* $\mathbf{x}_t$ *with* $\|\mathbf{x}_t\| \leq R$. *Let* $d \geq \left\lceil 18\frac{R^2T}{p}\log\frac{2T^2}{pc}\right\rceil$ *Suppose* U *is chosen uniformly at random from the set of* $d \times T$ *matrices with orthonormal columns. Let* $\textsc{Grad}$ *be an probability-p zero chain and let* $\textsc{Grad}_U(\mathbf{x}, \mathbf{z}) = U\textsc{Grad}(U^\top \mathbf{x}, \mathbf{z})$. *Let* $\mathbf{x}_1, \mathbf{x}_2, \ldots$ *be the iterates of* A *when provided the stochastic gradient oracle* $\textsc{Grad}_U$. *Then with probability at least* $1 - c$ *(over the randomness of* U, *the oracle, and also the seed* s *of* A*):*

$$prog_{1/4}(U^\top \mathbf{x}_t) < T \ \ \textit{for all } t \leq \frac{T - \log(2/c)}{2p} \ .$$

F.2. Defining the "Hard" Instance

Now, we for the first time diverge from the construction of Arjevani et al. (2019) (albeit only slightly). Their construction uses a "shrinking function" $\rho_{R,d} : \mathbb{R}^d \to \mathbb{R}^d$ given by $\rho_{R,d}(\mathbf{x}) = \frac{\|\mathbf{x}\|}{\sqrt{1+\|\mathbf{x}\|^2/R^2}}$ as well as an additional quadratic term to overcome the limitation of bounded iterates. We cannot tolerate the non-Lipschitz quadratic term, so we replace it with a Lipschitz version $q_{B,d}(\mathbf{x}) = \mathbf{x}^\top \rho_{B,d}(\mathbf{x})$. Intuitively, $q_{B,d}$ behaves like $\|\mathbf{x}\|^2$ for small enough $\mathbf{x}$, but behaves like $\|\mathbf{x}\|$ for large $\|\mathbf{x}\|$. This pre-processed function is defined by:

$$\begin{aligned} \hat{F}_{T,U}(\mathbf{x}) &= F_{T,U}(\rho_{R,d}(\mathbf{x})) + \eta q_{B,T}(\rho_{R,T}^{-1}(U^\top \rho_{R,d}(\mathbf{x}))) \\ &= F_T(U^\top \rho_{R,d}(\mathbf{x})) + \eta q_{B,T}(\rho_{R,T}^{-1}(U^\top \rho_{R,d}(\mathbf{x}))) \\ &= F_T(U^\top \rho_{R,d}(\mathbf{x})) + \eta q_{B,d}(\mathbf{x}) \ . \end{aligned}$$

To see the last line above, note that since $q_{B,d}$ and $\rho_{R,d}$ are rotationally symmetric, and U has orthonormal columns, we have $U^\top \rho_{R,d}(\mathbf{x}) = \rho_{R,T}(U^\top \mathbf{x})$ and $q_{B,T}(U^\top \mathbf{x}) = q_{B,d}(\mathbf{x})$. Thus, we have $q_{B,T}(\rho_{R,T}^{-1}(U^\top \rho_{R,d}(\mathbf{x}))) = q_{B,d}(\mathbf{x})$. The stochastic gradient oracle associated with $\hat{F}_{T,U}(\mathbf{x})$ is

$$\begin{aligned} \widehat{\textsc{Grad}}_{T,U}(\mathbf{x}, \mathbf{z}) &= J[\rho_{R,d}](\mathbf{x})^\top U\textsc{Grad}_T(U^\top \rho_{R,d}(\mathbf{x}), \mathbf{z}) + \eta\nabla q_{B,d}(\mathbf{x}) \\ &= J[\rho_{R,d}](\mathbf{x})^\top U\textsc{Grad}_T(U^\top \rho_{R,d}(\mathbf{x}), \mathbf{z}) \\ &\qquad + \eta J[\rho_{R,d}](\mathbf{x})^\top U^\top J[\rho_{R,T}^{-1}](U^\top \rho_{R,d}(\mathbf{x}))^\top \nabla q_{B,T}(\rho_{R,T}^{-1}(U^\top \rho_{R,d}(\mathbf{x}))) \ . \end{aligned}$$

where $J[f](\mathbf{x})$ indicates the Jacobian of the function f evaluated at $(\mathbf{x})$.

A description of the relevant properties of q_B is provided in Section F.3.

Next we produce a variant on Lemma 6 from Arjevani et al. (2019). This is the most delicate part of our alteration, although the proof is still almost identical to that of Arjevani et al. (2019).

Lemma 25 (variant on Lemma 6 of (Arjevani et al., 2019))**.** *Let $R = B = 10G_0\sqrt{T}$. Let $\eta = 1/10$ and $c \in (0,1)$ and $p \in (0,1)$ and $T \in \mathbb{N}$. Set $d = \lceil 18\frac{R^2T}{p}\log\frac{2T^2}{pc}\rceil$ and let U be sampled uniformly from the set of $d \times T$ matrices with orthonormal columns. Define $\hat{F}_{T,U}$ and $\hat{\textsc{Grad}}_{U,T}$ as above. Suppose $A \in \mathrm{A}_{rand}$ and let $\mathbf{x}_1, \mathbf{x}_2, \ldots$ be the iterates of A when provided with $\hat{\textsc{Grad}}_{U,T}$ as input. Then with probability at least $1-c$:*

$$\|\nabla \hat{F}_{T,U}(\mathbf{x}_t)\| \geq 1/2 \quad \textit{for all } t \leq \frac{T - \log(2/c)}{2p} \ .$$

Proof. First, observe that $\textsc{Grad}_T$ is a probability p zero-chain. Next, q_B is rotationally symmetric and $\rho_R(\mathbf{x}) \propto \mathcal{X}$, we see that $\nabla q_B(\mathbf{x}) \propto \mathbf{x}$ so that $\mathrm{prog}_0(J[\rho_{R,T}^{-1}](\mathbf{x})^\top \nabla q_{B,T}(\rho_{R,T}^{-1}(\mathbf{x}))) = \mathrm{prog}_0(\mathbf{x})$. Thus, the oracle

$$\widetilde{\textsc{Grad}}_T(\mathbf{x}, \mathbf{z}) = \textsc{Grad}_T(\mathbf{x}, \mathbf{z}) + \eta J[\rho_{B,T}^{-1}](\mathbf{x})^\top \nabla q_{B,T}(\rho_{R,T}^{-1}(\mathbf{x}))$$

is a probability-p zero-chain. We define the rotated oracle:

$$\begin{aligned}\widetilde{\textsc{Grad}}_{U,T}(\mathbf{y}, \mathbf{z}) &= U\widetilde{\textsc{Grad}}_T(U^\top \mathbf{y}, \mathbf{z}) \\ &= U\textsc{Grad}_T(U^\top \mathbf{y}, \mathbf{z}) + \eta U J[\rho_{B,T}^{-1}](U^\top \mathbf{y})^\top \nabla q_{B,T}(\rho_{R,T}^{-1}(U^\top \mathbf{y})) \ .\end{aligned}$$

Following (Arjevani et al., 2019), we now define $\mathbf{y}_i = \rho_R(\mathbf{x}_i)$. Notice that:

$$\begin{aligned}\widetilde{\textsc{Grad}}_{U,T}(\mathbf{y}, \mathbf{z}) &= J[\rho_{R,d}](\mathbf{x})^\top \widehat{\textsc{Grad}}_{T,U}(\mathbf{x}, \mathbf{z}) \\ &= J[\rho_{R,d}](\rho_{R,d}^{-1}(\mathbf{y}))]^\top \widehat{\textsc{Grad}}_{T,U}(\mathbf{x}, \mathbf{z}) \ .\end{aligned}$$

Therefore, we can view $\mathbf{y}_1, \ldots, \mathbf{y}_T$ as the iterates of some *different* algorithm $A^y \in \mathrm{A}_{\mathrm{rand}}$ applied to the oracle $\tilde{\textsc{Grad}}_{T,U}$ (since they are computed by measurable functions of the oracle outputs and previous iterates). However, since $\rho_R(\|bx\|) \leq R$, $\|\mathbf{y}_i\| \leq R$ for all i. Thus, since $\widetilde{\textsc{Grad}}_T$ is a probability-p zero-chain, by Lemma 24 we have that so long as $d \geq \lceil 18\frac{R^2T}{p}\log\frac{2T^2}{pc}\rceil$, then with probability at least $1-c$, for all $i \leq (T - \log(2/c))/2p$,

$$\mathrm{prog}_{1/4}(U^\top y_i) < T \ .$$

Now, our goal is to show that $\|\nabla F(\mathbf{x}_i)\| \geq 1/2$. We consider two cases, either $\|\mathbf{x}_i\| > R/2$ or not.

First, observe that for $\mathbf{x}_i$ with $\|\mathbf{x}_i\| > R/2$, we have:

$$\begin{aligned}\|\nabla \hat{F}_{T,U}(\mathbf{x}_i)\| &\geq \eta\|\nabla q_B(\|\mathbf{x}_i\|)\| - \|J[\rho_R](\mathbf{x}_i)\|_{\mathrm{op}}\|\nabla \hat{F}(\mathbf{y}_i)\| \\ &\geq \eta\frac{\|x_i\|}{\sqrt{1 + \|x_i\|^2/B^2}} - G_0\sqrt{T} \\ &\geq 3\eta B - G_0\sqrt{T}\end{aligned}$$

Recalling $B = R = 10G_0\sqrt{T}$ and $\eta = 1/10$:

$$= 2G_0\sqrt{T}$$

Recalling $G_0 = 23$:

$$\geq 1/2 \ .$$

Now, suppose $\|\mathbf{x}_i\| \le R/2$. Then, let us set $j = \text{prog}_1(U^\top \mathbf{y}_i) + 1 \le T$ (the inequality follows since $\text{prog}_1 \le \text{prog}_{1/4}$). Then, if $\mathbf{u}^j$ indicates the jth row of u, Lemma 22 implies:

$$|\langle \mathbf{u}^j, \mathbf{y}_i\rangle| < 1,$$
$$|\langle \mathbf{u}^j, \nabla F_{T,U}(\mathbf{y}_i)\rangle| \ge 1\ .$$

Next, by direct calculation we have $J[\rho_R](\mathbf{x}_i) = \frac{I - \rho_R(\mathbf{x}_i)\rho_R(\mathbf{x}_i)^\top/R^2}{\sqrt{1+\|\mathbf{x}_i\|^2/R^2}} = \frac{I - \mathbf{y}_i\mathbf{y}_i^\top/R^2}{\sqrt{1+\|\mathbf{x}_i\|^2/R^2}}$ so that:

$$\begin{aligned}\langle \mathbf{u}^j, \nabla \hat{F}_{T,U}(\mathbf{x}_i)\rangle &= \langle u^j J[\rho_R](\mathbf{x}_i)^\top \nabla F_{T,U}(\mathbf{y}_i)\rangle + \eta\langle \mathbf{u}^j, \nabla q_B(\mathbf{x}_i)\rangle \\ &= \frac{\langle \mathbf{u}^j, \nabla F_{T,U}(\mathbf{y}_i)\rangle}{\sqrt{1+\|\mathbf{x}_i\|^2/R^2}} - \frac{\langle \mathbf{u}^j, \mathbf{y}_i\rangle\langle \mathbf{y}_i, \nabla F_{T,U}(\mathbf{y}_i)\rangle/R^2}{\sqrt{1+\|\mathbf{x}_i\|^2/R^2}} + \eta\langle \mathbf{u}^j, \nabla q_B(\mathbf{x}_i)\rangle\ .\end{aligned}$$

Now, by Proposition 29, we have $\nabla q_B(\mathbf{x}_i) = \left(2 - \frac{\|\mathbf{y}_i\|^2}{B^2}\right)\mathbf{y}_i =$. So, (with $R = B$):

$$\langle \mathbf{u}^j, \nabla \hat{F}_{T,U}(\mathbf{x}_i)\rangle = \frac{\langle \mathbf{u}^j, \nabla F_{T,U}(\mathbf{y}_i)\rangle}{\sqrt{1+\|\mathbf{x}_i\|^2/R^2}} - \frac{\langle \mathbf{u}^j, \mathbf{y}_i\rangle\langle \mathbf{y}_i, \nabla F_{T,U}(\mathbf{y}_i)\rangle/R^2}{\sqrt{1+\|\mathbf{x}_i\|^2/R^2}} + \eta\left(2 - \frac{\|\mathbf{y}_i\|^2}{B^2}\right)\langle \mathbf{u}^j, \mathbf{y}_i\rangle$$

Observing that $\|y_i\| \le \|x_i\| \le R/2$ and $|\langle \mathbf{u}^j, \mathbf{y}_i\rangle| < 1$:

$$|\langle \mathbf{u}^j, \nabla \hat{F}_{T,U}(\mathbf{x}_i)\rangle| \ge \frac{1}{\sqrt{1+\|\mathbf{x}_i\|^2/R^2}} - \frac{\|\nabla F_{T,U}(\mathbf{y}_i)\|}{2R} - 2\eta$$

Using $|\langle \mathbf{u}^j, \nabla F_{T,U}(\mathbf{y}_i)\rangle| \ge 1$ and $\|\nabla F(\mathbf{y}_i)\| \le G_0\sqrt{T}$:

$$\ge \frac{2}{\sqrt{5}} - \frac{G_0\sqrt{T}}{2R} - 2\eta$$

With $R = 10G_0\sqrt{T}$ and $\eta = 1/10$:

$$\begin{aligned}&= \frac{2}{\sqrt{5}} - \frac{1}{20} - \frac{1}{5} \\ &> 1/2\ .\end{aligned}$$

□

With the settings of parameters specified in Lemma 25, we have:

Lemma 26 (variation on Lemma 7 in Arjevani et al. (2019))**.** *With the settings of R, B, η in Lemma 25, the function $\hat{F}_{T,U}$ satisfies:*

1. $\hat{F}_{T,U}(0) - \inf \hat{F}_{T,U}(\mathbf{x}) \le \gamma_0 T = 12T$
2. $\|\hat{F}_{T,U}(\mathbf{x})\| \le G_0\sqrt{T} + 3\eta B \le 92\sqrt{T}$ *for all* $\mathbf{x}$.
3. $\nabla \hat{F}_U(\mathbf{x})$ *is* $H_0 + 3 + 8\eta \le 156$*-Lipschitz.*
4. $\|\widehat{\text{GRAD}}_{T,U}(\mathbf{x}, \mathbf{z})\| \le \frac{G_0}{p} + G_0\sqrt{T} + 3\eta B \le \frac{23}{p} + 92\sqrt{T}$ *with probability 1.*
5. $\text{GRAD}_{T,U}$ *has variance at most* $\frac{G_0^2}{p} \le \frac{23^2}{p}$

Proof. 1. This property follows immediately from the fact that $F_T(0) - \inf F_T(\mathbf{x}) \le \gamma_0 T$.

2. Since ρ_R is 1-Lipschitz for all R and q_B is $3B$-Lipschitz (see Proposition 29), $\hat{F}_{T,U}(\mathbf{x})$ is $G_0\sqrt{T} + 3\eta B$-Lipschitz.
3. By assumption, $R \ge \max(H_0, 1)$. Thus, by Arjevani et al. (2019, Lemma 16), $\nabla F_T(\rho_R(\mathbf{x}))$ is $H_0 + 3$-Lipschitz and so $\nabla \hat{F}_{T,U}$ is $H_0 + 3 + 8\eta$-Lipschitz by Proposition 29.

4. Since $\|O_T\| \leq \frac{G_0}{p} + G_0\sqrt{T}$, and $J[\rho_R](\mathbf{x})^\top U$ has operator norm at most 1, the bound follows.
5. Just as in the previous part, since $\textsc{Grad}_T$ has variance G_0^2/p and $J[\rho_R](\mathbf{x})^\top U$ has operator norm at most 1, the bound follows.

□

Now, we are finally in a position to prove:

Theorem 27. *Given any γ, H, ϵ, and σ such that $\frac{H\gamma}{12\cdot 156\cdot 4\epsilon^2} \geq 2$, there exists a distribution over functions F and stochastic first-order oracles* GRAD *such that with probability 1, F is H-smooth, $F(0) - \inf F(\mathbf{x}) \leq \gamma$, F is $3\sqrt{H\gamma}$-Lipschitz and* GRAD *has variance σ^2, and for any algorithm in A_{rand}, with probability at least $1-c$, when provided a randomly selected* GRAD*, A requires at least $\Omega\left(\frac{\gamma H\sigma^2}{\epsilon^4}\right)$ iterations to output a point $\mathbf{x}$ with $\mathbb{E}[\|\nabla F(\mathbf{x})\|] \leq \epsilon$.*

Proof. From Lemma 26 and Lemma 25, we have a distribution over functions F and first-order oracles such that with probability 1, F is $G_0\sqrt{T}+3G_0\sqrt{T} \leq 92\sqrt{T}$ Lipschitz, F is $H_0+3+4\eta \leq 156$-smoooth, $F(0)-\inf F(\mathbf{x}) \leq \gamma_0 T = 12T$, GRAD has variance at most $G_0^2/p = 23^2/p$, and with probability at least $1-c$,

$$\|\nabla F(\mathbf{x}_t)\| \geq 1/2 \quad \text{for all } t \leq \frac{T-\log(2/c)}{2p} .$$

Now, set $\lambda = \frac{156}{H}\cdot 2\epsilon$, $T = \lfloor \frac{H\gamma}{12\cdot 156\cdot (2\epsilon)^2} \rfloor$ and $p = \min((2\cdot 23\cdot \epsilon)^2/\sigma^2, 1)$. Then, define

$$F_\lambda(\mathbf{x}) = \frac{H\lambda^2}{156} F(\mathbf{x}/\lambda) .$$

Then F_λ is $\frac{H\lambda^2}{H_0}\cdot\frac{1}{\lambda^2}\cdot 156 = H$- smooth, $F_\lambda(0) - \inf_{\mathbf{x}} F_\lambda(\mathbf{x}) \leq 12\cdot T\cdot \frac{H\lambda^2}{156} \leq \gamma$, and F_λ is $92\sqrt{T}\frac{H\lambda}{156} \leq 3\sqrt{H\gamma}$- Lipschitz. We can construct an oracle $\textsc{Grad}_\lambda$ from GRAD by:

$$\textsc{Grad}_\lambda(\mathbf{x},\mathbf{z}) = \frac{H\lambda}{156}\textsc{Grad}(\mathbf{x}/\lambda, \mathbf{z}) .$$

so that

$$\mathbb{E}[\|\textsc{Grad}_\lambda(\mathbf{x},\mathbf{z}) - \nabla F_\lambda(\mathbf{x})\|^2] \leq \frac{H^2\lambda^2}{156^2}\cdot\frac{23^2}{p} = \sigma^2 .$$

Further, since an oracle for F_λ can be constructed from the oracle for F, if we run A on F_λ, with probability at least $1-c$,

$$\|\nabla F_\lambda(\mathbf{x}_t)\| = \frac{H\lambda}{156}\|\nabla F(\mathbf{x}_t)\| \geq \epsilon \quad \text{for all } t \leq \frac{T-\log(2/c)}{2p} \leq 3\cdot 10^{-7}\cdot\frac{H\gamma\sigma^2}{\epsilon^4} - 5\cdot 10^{-3}\cdot\frac{\log(2/c)}{\epsilon^2} .$$

Thus, there exists a constant K such that

$$\mathbb{E}[\|\nabla F_\lambda(\mathbf{x}_t)\|] \geq \epsilon \quad \text{for all } t \leq K\frac{H\gamma\sigma^2}{\epsilon^4} .$$

□

From this result, we have our main lower bound (the formal version of Theorem 18):

Theorem 28. *For any δ, ϵ, γ, $G \geq \frac{3\sqrt{2\epsilon\gamma}}{\sqrt{\delta}}$, there is a distribution over G-Lipschitz C^∞ functions F with $F(0)-\inf F(\mathbf{x}) \leq \gamma$ and stochastic gradient oracles* GRAD *with $\mathbb{E}[\|\textsc{Grad}(\mathbf{x},\mathbf{z})\|^2] \leq G^2$ such that for any algorithm $A \in \mathrm{A}_{rand}$, if A is provided as input a randomly selected oracle* GRAD*, A will require $\Omega(G^2\gamma/\delta\epsilon^3)$ iterations to identify a point x with $\mathbb{E}[\|\nabla F(\mathbf{x})\|_\delta] \leq \epsilon$.*

Proof. From Theorem 27, for any H, ϵ', γ, and σ we have a distrubution over C^∞ functions F and oracles GRAD such that F is H-smooth, $3\sqrt{H\gamma}$-Lischitz and $F(0) - \inf F(\mathbf{x}) \leq \gamma$ and GRAD has variance σ^2 such that A requires $\Omega(H\gamma\sigma^2/\epsilon'^4)$ iterations to output a point $\mathbf{x}$ such that $\mathbb{E}[\|\nabla F(\mathbf{x})\|] \leq \epsilon'$. Set $\sigma = G$, $H = \epsilon/\delta$ and $\epsilon' = 2\epsilon$. Then, we see that GRAD has variance $G^2/2$, and F is $3\sqrt{H\gamma} = \frac{3\sqrt{\epsilon\gamma}}{\sqrt{\delta}} \leq G/\sqrt{2}$-Lipschitz so that $\mathbb{E}[\|\textsc{Grad}(\mathbf{x},\mathbf{z})\|^2] \leq G^2$. Further by Proposition!14, if $\|\nabla F(\mathbf{x})\|_\delta \leq \epsilon$, then $\|\nabla F(\mathbf{x})\| \leq \epsilon + H\delta = 2\epsilon$. Therefore, since A cannot output a point $\mathbf{x}$ with $\mathbb{E}[\|\nabla F(\mathbf{x})\|] \leq \epsilon' = 2\epsilon$ in less than $\Omega(H\gamma\sigma^2/\epsilon'^4)$ iterations, we see that A also cannot output a point $\mathbf{x}$ with $\mathbb{E}[\|\nabla F(\mathbf{x})\|_\delta] \leq \epsilon$ in less than $\Omega(H\gamma\sigma^2/\epsilon'^4) = \Omega(\gamma G^2/\epsilon^3\delta)$ iterations. □

F.3. Definition and Properties of q_B

Consider the function $q_{B,d} : \mathbb{R}^d \to \mathbb{R}$ defined by

$$q_{B,d}(\mathbf{x}) = \frac{\|\mathbf{x}\|^2}{\sqrt{1+\|\mathbf{x}\|^2/B^2}} = \mathbf{x}^\top \rho_{B,d}(\mathbf{x}) \ .$$

This function has the following properties, all of which follow from direct calculuation:

Proposition 29. *$q_{B,d}$ satisfies:*

1.
$$\nabla q_{B,d}(\mathbf{x}) = \frac{2\mathbf{x}}{\sqrt{1+\|\mathbf{x}\|^2/B^2}} - \frac{\mathbf{x}\|\mathbf{x}\|^2}{B^2(1+\|\mathbf{x}\|^2/B^2)^{3/2}} = \left(2 + \frac{\|\rho_B(\mathbf{x})\|^2}{B^2}\right)\rho_B(\mathbf{x}) \ .$$

2.
$$\nabla^2 q_{B,d}(\mathbf{x}) = \frac{1}{\sqrt{1+\|\mathbf{x}\|^2/B^2}}\left(2I - \frac{3\mathbf{x}\mathbf{x}^\top}{B^2(1+\|\mathbf{x}\|^2/B^2)} - \frac{\|\mathbf{x}\|^2 I}{B^2(1+\|\mathbf{x}\|^2/B^2)} + \frac{2\|\mathbf{x}\|^2\mathbf{x}\mathbf{x}^\top}{B^4(1+\|\mathbf{x}\|^2/B^2)^2}\right) \ .$$

3.
$$\frac{\|\mathbf{x}\|}{\sqrt{1+\|\mathbf{x}\|^2/B^2}} \le \|\nabla q_{B,d}(\mathbf{x})\| \le \frac{3\|\mathbf{x}\|}{\sqrt{1+\|\mathbf{x}\|^2/B^2}} \le 3B \ .$$

4.
$$\|\nabla^2 q_{B,d}(\mathbf{x})\|_{op} \le \frac{8}{\sqrt{1+\|\mathbf{x}\|^2/B^2}} \le 8 \ .$$

G. Proof of Theorem 13

First, we state and prove a theorem analogous to Theorem 8.

Theorem 30. *Assume $F : \mathbb{R}^d \to \mathbb{R}$ is well-behaved. In Algorithm 1, set s_n to be a random variable sampled uniformly from $[0,1]$. Set $T, K \in \mathbb{N}$ and $M = KT$. For $i = 1, \dots, d$, set $u_i^k = -D_\infty \frac{\sum_{t=1}^T \frac{\partial F\left(\mathbf{w}_t^k\right)}{\partial x_i}}{\left|\sum_{t=1}^T \frac{\partial F\left(\mathbf{w}_t^k\right)}{\partial x_i}\right|}$ for some $D_\infty > 0$. Finally, suppose $Var(g_{n,i}) \le \sigma_i^2$ for $i = 1, \dots, d$. Then, we have*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^K \left\|\frac{1}{T}\sum_{t=1}^T \nabla F(\mathbf{w}_t^k)\right\|_1\right] \le \frac{F(\mathbf{x}_0) - F^\star}{D_\infty M} + \frac{\mathbb{E}[R_T(\mathbf{u}^1, \dots, \mathbf{u}^K)]}{D_\infty M} + \frac{D_\infty \sum_{i=1}^d \sigma_i}{\sqrt{T}} \ .$$

Proof. In Theorem 7, set $\mathbf{u}_n$ to be equal to $\mathbf{u}^1$ for the first T iterations, $\mathbf{u}^2$ for the second T iterations and so on. In other words, $\mathbf{u}_n = \mathbf{u}^{mod(n,T)+1}$ for $n = 1, \dots, N$.

From Theorem 7, we have

$$\mathbb{E}[F(\mathbf{x}_M)] = F(\mathbf{x}_0) + \mathbb{E}[R_T(\mathbf{u}^1, \dots, \mathbf{u}^K)] + \mathbb{E}\left[\sum_{n=1}^M \langle \mathbf{g}_n, \mathbf{u}_n\rangle\right] \ .$$

Now, since $u_{k,i} = -D_\infty \frac{\sum_{t=1}^T \frac{\partial F\left(\mathbf{w}_t^k\right)}{\partial x_i}}{\left|\sum_{t=1}^T \frac{\partial F\left(\mathbf{w}_t^k\right)}{\partial x_i}\right|}$, $\mathbb{E}[\mathbf{g}_n] = \nabla F(\mathbf{w}_n)$, and $\text{Var}(g_{n,i}) \leq \sigma_i^2$ for $i = 1, \dots, d$, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{n=1}^M \langle \mathbf{g}_n, \mathbf{u}_n \rangle\right] &\leq \mathbb{E}\left[\sum_{k=1}^K \left\langle \sum_{t=1}^T \nabla F(\mathbf{w}_t^k), \mathbf{u}^k \right\rangle + D_\infty \sum_{k=1}^K \left\| \sum_{t=1}^T (\nabla F(\mathbf{w}_t^k) - \mathbf{g}_n) \right\|_1\right] \\
&\leq \mathbb{E}\left[\sum_{k=1}^K \left\langle \sum_{t=1}^T \nabla F(\mathbf{w}_t^k), \mathbf{u}^k \right\rangle\right] + D_\infty K\sqrt{T} \sum_{i=1}^d \sigma_i \\
&= \mathbb{E}\left[-\sum_{k=1}^K D_\infty T \left\| \frac{1}{T} \sum_{t=1}^T \nabla F\left(\mathbf{w}_t^k\right) \right\|_1\right] + D_\infty K\sqrt{T} \sum_{i=1}^d \sigma_i \,.
\end{aligned}$$

Putting this all together, we have

$$F^\star \leq \mathbb{E}[F(\mathbf{x}_N)] \leq F(\mathbf{x}_0) + \mathbb{E}[R_T(\mathbf{u}^1, \dots, \mathbf{u}^k)] + D_\infty K\sqrt{T} \sum_{i=1}^d \sigma_i - D_\infty T \sum_{k=1}^K \mathbb{E}\left[\left\| \frac{1}{T} \sum_{t=1}^T \nabla F\left(\mathbf{w}_t^k\right) \right\|_1\right] \,.$$

Dividing by $KTD_\infty = D_\infty M$ and reordering, we have the stated bound. □

We can now prove Theorem 13.

Proof of Theorem 13. Since $\mathcal{A}$ guarantees $\|\mathbf{\Delta}_n\|_\infty \leq D_\infty$, for all $n < n' \leq T + n - 1$, we have

$$\begin{aligned}
\|\mathbf{w}_n - \mathbf{w}_{n'}\|_\infty &= \|\mathbf{x}_n - (1 - s_n)\mathbf{\Delta}_n - \mathbf{x}_{n'-1} + s_{n'}\mathbf{\Delta}_{n'}\|_\infty \\
&\leq \left\| \sum_{i=n+1}^{n'-1} \mathbf{\Delta}_i \right\|_\infty + \|\mathbf{\Delta}_n\|_\infty + \|\mathbf{\Delta}_{n'}\|_\infty \\
&\leq D_\infty((n' - 1) - (n + 1) + 1) + 2D_\infty \\
&= D_\infty(n' - n + 1) \\
&\leq D_\infty T \,.
\end{aligned}$$

Therefore, we clearly have $\|\mathbf{w}_t^k - \overline{\mathbf{w}}^k\|_\infty \leq D_\infty T = \delta$.

Note that from the choice of K and T we have $M = KT \geq N - T \geq N/2$. Now, observe that $\text{Var}(g_{n,i}) \leq \mathbb{E}[g_{n_i}^2] \leq G_i^2$. Thus, applying Theorem 30 in concert with the additional assumption $\mathbb{E}[R_T(\mathbf{u}^1, \dots, \mathbf{u}^K)] \leq D_\infty K\sqrt{T} \sum_{i=1}^d G_i$, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{K} \sum_{k=1}^K \left\| \frac{1}{T} \sum_{t=1}^T \nabla F(\mathbf{w}_t^k) \right\|_1\right] &\leq 2\frac{F(\mathbf{x}_0) - F^\star}{D_\infty N} + 2\frac{KD_\infty\sqrt{T}\sum_{i=1}^d G_i}{D_\infty N} + \frac{\sum_{i=1}^d G_i}{\sqrt{T}} \\
&= \frac{2T(F(\mathbf{x}_0) - F^\star)}{\delta N} + \frac{3\sum_{i=1}^d G_i}{\sqrt{T}} \\
&\leq \max\left(\frac{5(\sum_{i=1}^d G_i)^{2/3}(F(\mathbf{x}_0) - F^\star)^{1/3}}{(N\delta)^{1/3}}, \frac{6\sum_{i=1}^d G_i}{\sqrt{N}}\right) + \frac{2(F(\mathbf{x}_0) - F^\star)}{\delta N},
\end{aligned}$$

where the last inequality is due to the choice of T.

Now to conclude, observe that $\|\mathbf{w}_t^k - \overline{\mathbf{w}}^k\|_\infty \leq \delta$ for all t and k, and also that $\overline{\mathbf{w}}^k = \frac{1}{T}\sum_{t=1}^T \mathbf{w}_t^k$. Therefore $S = \{\mathbf{w}_1^k, \dots, \mathbf{w}_T^k\}$ satisfies the conditions in the infimum in Definition 12 so that $\|\nabla F(\overline{\mathbf{w}}^k)\|_{1,\delta} \leq \left\|\frac{1}{T}\sum_{t=1}^T \nabla F(\mathbf{w}_t^k)\right\|_1$. □

H. Directional Derivative Setting

In the main text, our algorithms make use of a stochastic gradient oracle. However, the prior work of (Zhang et al., 2020b) instead considers a stochastic *directional* gradient oracle. This is a less common setup, and other works (e.g., (Davis et al., 2021)) have also taken our route of tackling non-smooth optimization via an oracle that returns gradients at points of differentiability.

Nevertheless, all our results extend easily to the exact setting of (Zhang et al., 2020b) in which F is Lipschitz and directionally differentiable and we have access to a stochastic directional gradient oracle rather than a stochastic gradient oracle. To quantify this setting, we need a bit more notation which we copy directly from (Zhang et al., 2020b) below:

First, from (Clarke, 1990) and (Zhang et al., 2020b), the *generalized directional derivative* of a function F in a direction $\mathbf{d}$ is

$$F^\circ(\mathbf{x}, \mathbf{d}) = \limsup_{\mathbf{y} \to \mathbf{x}\ t \downarrow 0} \frac{f(\mathbf{y} + t\mathbf{d}) - f(\mathbf{y})}{t} . \tag{4}$$

Further, the *generalized gradient* is the set

$$\partial F(\mathbf{x}) = \{\mathbf{g} : \ \langle \mathbf{g}, \mathbf{d} \rangle \leq \langle F^\circ(\mathbf{x}, \mathbf{d}), \mathbf{d} \rangle \text{ for all } \mathbf{d}\} .$$

Finally, $F : \mathbb{R}^d \to \mathbb{R}$ is *Hadamard directionally differentiable* in the direction $\mathbf{v} \in \mathbb{R}^d$ if for any function $\psi : \mathbb{R}_+ \to \mathbb{R}^d$ such that $\lim_{t \to 0} \frac{\psi(t) - \psi(0)}{t} = \mathbf{v}$ and $\psi(0) = \mathbf{x}$, the following limit exists:

$$\lim_{t \to 0} \frac{F(\psi(t)) - F(\mathbf{x})}{t} .$$

If F is Hadamard directionally differentiable, then the above limit is denoted $F'(\mathbf{x}, \mathbf{v})$. When F is Hadamard directionally differentiable for all $\mathbf{x}$ and $\mathbf{v}$, then we say simply that F is directionally differentiable.

With these definitions, a *stochastic directional oracle* for a Lipschitz, directionally differentiable, and bounded from below function F is an oracle $\textsc{Grad}(\mathbf{x}, \mathbf{v}, \mathbf{z})$ that outputs $\mathbf{g} \in \partial F(\mathbf{x})$ such that $\langle \mathbf{g}, \mathbf{v} \rangle = F'(\mathbf{x}, \mathbf{v})$. In this case, (Zhang et al., 2020b) shows (Lemma 3) that F satisfies an alternative notion of well-behavedness:

$$F(\mathbf{y}) - F(\mathbf{x}) = \int_0^1 \langle \mathbb{E}[\textsc{Grad}(\mathbf{x} + t(\mathbf{y} - \mathbf{x}), \mathbf{y} - \mathbf{x}, \mathbf{z})], \mathbf{y} - \mathbf{x} \rangle dt . \tag{5}$$

Next, we define:

Definition 31. *A point $\mathbf{x}$ is a (δ, ϵ) stationary point of F for the generalized gradient if there is a set of points S contained in the ball of radius δ centered at $\mathbf{x}$ such that for $\mathbf{y}$ selected uniformly at random from S, $\mathbb{E}[\mathbf{y}] = \mathbf{x}$ and for all $\mathbf{y}$ there is a choice of $\mathbf{g}_\mathbf{y} \in \partial F(\mathbf{y})$ such that $\| \mathbb{E}[\mathbf{g}_\mathbf{y}] \| \leq \epsilon$.*

Similarly, we have the definition:

Definition 32. *Given a point $\mathbf{x}$, and a number $\delta > 0$, define:*

$$\|\partial F(\mathbf{x})\|_\delta \triangleq \inf_{S \subset B(\mathbf{x}, \delta), \frac{1}{|S|} \sum_{\mathbf{y} \in S} \mathbf{y} = \mathbf{x}, \mathbf{g}_\mathbf{y} \in \partial F(\mathbf{y})} \left\| \frac{1}{|S|} \sum_{\mathbf{y} \in S} \mathbf{g}_\mathbf{y} \right\| .$$

In fact, whenever a locally Lipschitz function F is differentiable at a point $\mathbf{x}$, we have that $\nabla F(\mathbf{x}) \in \partial F(\mathbf{x})$, so that $\|\partial F(\mathbf{x})\|_\delta \leq \|\nabla F(\mathbf{x})\|_\delta$. Thus our results in the main text also bound $\|\partial F(\mathbf{x})\|_\delta$. However, while a gradient oracle is also directional derivative oracle, a directional derivative oracle is only guaranteed to be a gradient oracle if F is *continuously* differentiable at the queried point $\mathbf{x}$. This technical issue means that when we have access to a directional derivative oracle rather than a gradient oracle, we will instead only bound $\|\partial F(\mathbf{x})\|_\delta$ rather than $\|\nabla F(\mathbf{x})\|_\delta$.

Despite this technical complication, our overall strategy is essentially identical. The key observation is that the only time at which we used the properties of the gradient previously was when we invoked well-behavedness of F. When we have a directional derivative instead of the gradient, the alternative notion of well-behavedness in (5) will play an identical role.

Algorithm 4 Online-to-Non-Convex Conversion (directional derivative oracle version)

Input: Initial point $\mathbf{x}_0$, $K \in \mathbb{N}$, $T \in \mathbb{N}$, online learning algorithm $\mathcal{A}$, s_n for all n
Set $M = K \cdot T$
for $n = 1 \dots M$ **do**
  Get $\boldsymbol{\Delta}_n$ from $\mathcal{A}$
  Set $\mathbf{x}_n = \mathbf{x}_{n-1} + \boldsymbol{\Delta}_n$
  Set $\mathbf{w}_n = \mathbf{x}_{n-1} + s_n \boldsymbol{\Delta}_n$
  Sample random $\mathbf{z}_n$
  Generate directional derivative $\mathbf{g}_n = \textsc{Grad}(\mathbf{w}_n, \boldsymbol{\Delta}_n, \mathbf{z}_n)$
  Send $\mathbf{g}_n$ to $\mathcal{A}$ as gradient
end for
Set $\mathbf{w}_t^k = \mathbf{w}_{(k-1)T+t}$ for $k = 1, \dots, K$ and $t = 1, \dots, T$
Set $\overline{\mathbf{w}}^k = \frac{1}{T}\sum_{t=1}^T \mathbf{w}_t^k$ for $k = 1, \dots, K$
Return $\{\overline{\mathbf{w}}^1, \dots, \overline{\mathbf{w}}^K\}$

Thus, our approach is simply to replace the call to $\textsc{Grad}(\mathbf{w}_n, \mathbf{z}_n)$ in Algorithm 1 with a call instead to $\textsc{Grad}(\mathbf{w}_n, \boldsymbol{\Delta}_n, \mathbf{z}_n)$ (see Algorithm 4). With this change, all of our analysis in the main text applies almost without modification. Essentially, we only need to change notation in a few places to reflect the updated definitions.

To begin this notational update, the counterpart to Theorem 7 is:

Theorem 33. *Suppose F is Lipschitz and directionally differentiable. With the notation in Algorithm 4, if we let s_n be independent random variables uniformly distributed in $[0,1]$, then for any sequence of vectors $\mathbf{u}_1, \dots, \mathbf{u}_N$, if we have the* equality*:*

$$\mathbb{E}[F(\mathbf{x}_M)] = F(\mathbf{x}_0) + \mathbb{E}\left[\sum_{n=1}^M \langle \mathbf{g}_n, \boldsymbol{\Delta}_n - \mathbf{u}_n\rangle\right] + \mathbb{E}\left[\sum_{n=1}^M \langle \mathbf{g}_n, \mathbf{u}_n\rangle\right] .$$

Proof.

$$\begin{aligned} F(\mathbf{x}_n) - F(\mathbf{x}_{n-1}) &= \int_0^1 \langle \mathbb{E}[\textsc{Grad}(\mathbf{x}_{n-1} + s(\mathbf{x}_n - \mathbf{x}_{n-1}), \mathbf{x}_n - \mathbf{x}_{n-1}, \mathbf{z}_n)], \mathbf{x}_n - \mathbf{x}_{n-1}\rangle \, \mathrm{d}s \\ &= \mathbb{E}[\langle \mathbf{g}_n, \boldsymbol{\Delta}_n\rangle] \\ &= \mathbb{E}[\langle \mathbf{g}_n, \boldsymbol{\Delta}_n - \mathbf{u}_n\rangle + \langle \mathbf{g}_n, \mathbf{u}_n\rangle] . \end{aligned}$$

Where in the second line we have used the definition $\mathbf{g}_n = \textsc{Grad}(\mathbf{x}_{n-1} + s_n(\mathbf{x}_n - \mathbf{x}_{n-1}), \mathbf{x}_n - \mathbf{x}_{n-1}, \mathbf{z}_n)$, the assumption that s_n is uniform on $[0,1]$, and Fubini theorem (as $\textsc{Grad}$ is bounded by Lipschitzness of F). Now, sum over n and telescope to obtain the stated bound.

□

Next, we have the following analog of Theorem 8:

Theorem 34. *With the notation in Algorithm 4, set s_n to be a random variable sampled uniformly from $[0,1]$. Set $T, K \in \mathbb{N}$ and $M = KT$. Define $\boldsymbol{\nabla}_t^k = \mathbb{E}[\mathbf{g}_{(k-1)T+t}]$. Define $\mathbf{u}^k = -D\frac{\sum_{t=1}^T \boldsymbol{\nabla}_t^k}{\|\sum_{t=1}^T \boldsymbol{\nabla}_t^k\|}$ for some $D > 0$ for $k = 1, \dots, K$. Finally, suppose Var$(\mathbf{g}_n) = \sigma^2$. Then:*

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^K \left\|\frac{1}{T}\sum_{t=1}^T \boldsymbol{\nabla}_t^k\right\|\right] \leq \frac{F(\mathbf{x}_0) - F^\star}{DM} + \frac{\mathbb{E}[R_T(\mathbf{u}^1, \dots, \mathbf{u}^K)]}{DM} + \frac{\sigma}{\sqrt{T}} .$$

Proof. The proof is essentially identical to that of Theorem 8. In Theorem 33, set $\mathbf{u}_n$ to be equal to $\mathbf{u}^1$ for the first T iterations, $\mathbf{u}^2$ for the second T iterations and so on. In other words, $\mathbf{u}_n = \mathbf{u}^{mod(n,T)+1}$ for $n = 1, \dots, M$. So, we have

$$\mathbb{E}[F(\mathbf{x}_M)] = F(\mathbf{x}_0) + \mathbb{E}[R_T(\mathbf{u}^1, \dots, \mathbf{u}^K)] + \mathbb{E}\left[\sum_{n=1}^M \langle \mathbf{g}_n, \mathbf{u}_n\rangle\right] .$$

Now, since $\mathbf{u}^k = -D\frac{\sum_{t=1}^T \boldsymbol{\nabla}_t^k}{\|\sum_{t=1}^T \boldsymbol{\nabla}_t^k\|}$, and $\mathrm{Var}(\mathbf{g}_n) = \sigma^2$, we have

$$\begin{aligned}
\mathbb{E}\left[\sum_{n=1}^M \langle \mathbf{g}_n, \mathbf{u}_n\rangle\right] &\leq \mathbb{E}\left[\sum_{k=1}^K \left\langle \sum_{t=1}^T \boldsymbol{\nabla}_t^k, \mathbf{u}^k\right\rangle\right] + \mathbb{E}\left[D\sum_{k=1}^K \left\|\sum_{t=1}^T (\boldsymbol{\nabla}_t^k - \mathbf{g}_{(k-1)T+t})\right\|\right] \\
&\leq \mathbb{E}\left[\sum_{k=1}^K \left\langle \sum_{t=1}^T \boldsymbol{\nabla}_t^k, \mathbf{u}^k\right\rangle\right] + D\sigma K\sqrt{T} \\
&= \mathbb{E}\left[-\sum_{k=1}^K DT\left\|\frac{1}{T}\sum_{t=1}^T \boldsymbol{\nabla}_t^k\right\|\right] + D\sigma K\sqrt{T}\,.
\end{aligned}$$

Putting this all together, we have

$$F^\star \leq \mathbb{E}[F(\mathbf{x}_M)] \leq F(\mathbf{x}_0) + \mathbb{E}[R_T(\mathbf{u}^1, \ldots, \mathbf{u}^K)] + \sigma DK\sqrt{T} - DT\sum_{k=1}^K \mathbb{E}\left[\left\|\frac{1}{T}\sum_{t=1}^T \nabla F\left(\mathbf{w}_t^k\right)\right\|\right]\,.$$

Dividing by $KDT = DM$ and reordering, we have the stated bound. □

Finally, we instantiate Theorem 34 with online gradient descent to obtain the analog of Corollary 9. This result establishes that the online-to-batch conversion finds an (δ, ϵ) critical point in $O(1/\epsilon^3\delta)$ iterations, even when using a directional derivative oracle. Further, our lower bound construction makes use of continuously differentiable functions, for which the directional derivative oracle and the standard gradient oracle must coincide. Thus the $O(1/\epsilon^3\delta)$ complexity is optimal in this setting as well.

Corollary 35. *Suppose we have a budget of N gradient evaluations. Under the assumptions and notation of Theorem 34, suppose in addition $\mathbb{E}[\|\mathbf{g}_n\|^2] \leq G^2$ and that $\mathcal{A}$ guarantees $\|\boldsymbol{\Delta}_n\| \leq D$ for some user-specified D for all n and ensures the worst-case K-shifting regret bound $\mathbb{E}[R_T(\mathbf{u}^1, \ldots, \mathbf{u}^K)] \leq DGK\sqrt{T}$ for all $\|\mathbf{u}^k\| \leq D$ (e.g., as achieved by the OGD algorithm that is reset every T iterations). Let $\delta > 0$ be an arbitrary number. Set $D = \delta/T$, $T = \min(\lceil(\frac{GN\delta}{F(\mathbf{x}_0)-F^\star})^{2/3}\rceil, \frac{N}{2})$, and $K = \lfloor\frac{N}{T}\rfloor$. Then, for all k and t, $\|\overline{\mathbf{w}}^k - \mathbf{w}_t^k\| \leq \delta$.*

Moreover, we have the inequality

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^K \left\|\frac{1}{T}\sum_{t=1}^T \boldsymbol{\nabla}_t^k\right\|\right] \leq \frac{2(F(\mathbf{x}_0) - F^\star)}{\delta N} + \max\left(\frac{5G^{2/3}(F(\mathbf{x}_0) - F^\star)^{1/3}}{(N\delta)^{1/3}}, \frac{6G}{\sqrt{N}}\right),$$

which implies

$$\frac{1}{K}\sum_{t=1}^K \|\partial F(\overline{\mathbf{w}}^k)\|_\delta \leq \frac{2(F(\mathbf{x}_0) - F^\star)}{\delta N} + \max\left(\frac{5G^{2/3}(F(\mathbf{x}_0) - F^\star)^{1/3}}{(N\delta)^{1/3}}, \frac{6G}{\sqrt{N}}\right)\,.$$

Proof. Since $\mathcal{A}$ guarantees $\|\boldsymbol{\Delta}_n\| \leq D$, for all $n < n' \leq T + n - 1$, we have

$$\begin{aligned}
\|\mathbf{w}_n - \mathbf{w}_{n'}\| &= \|\mathbf{x}_n - (1 - s_n)\boldsymbol{\Delta}_n - \mathbf{x}_{n'-1} + s_{n'}\boldsymbol{\Delta}_{n'}\| \\
&\leq \left\|\sum_{i=n+1}^{n'-1} \boldsymbol{\Delta}_i\right\| + \|\boldsymbol{\Delta}_n\| + \|\boldsymbol{\Delta}_{n'}\| \\
&\leq D((n' - 1) - (n + 1) + 1) + 2D \\
&= D(n' - n + 1) \leq DT\,.
\end{aligned}$$

Therefore, we clearly have $\|\mathbf{w}_t^k - \overline{\mathbf{w}}^k\| \leq DT = \delta$.

Note that from the choice of K and T we have $M = KT \geq N - T \geq N/2$. So, for the second fact, notice that $\text{Var}(\mathbf{g}_n) \leq G^2$ for all n. Thus, applying Theorem 34 in concert with the additional assumption $\mathbb{E}[R_T(\mathbf{u}^1, \dots, \mathbf{u}^K)] \leq DGK\sqrt{T}$, we have:

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\nabla}_t^k\right\|\right] &\leq 2\frac{F(\mathbf{x}_0) - F^\star}{DN} + 2\frac{KDG\sqrt{T}}{DN} + \frac{G}{\sqrt{T}} \\
&\leq \frac{2T(F(\mathbf{x}_0) - F^\star)}{\delta N} + \frac{3G}{\sqrt{T}} \\
&\leq \max\left(\frac{5G^{2/3}(F(\mathbf{x}_0) - F^\star)^{1/3}}{(N\delta)^{1/3}}, \frac{6G}{\sqrt{N}}\right) + \frac{2(F(\mathbf{x}_0) - F^\star)}{\delta N},
\end{aligned}$$

where the last inequality is due to the choice of T.

Finally, observe that $\|\mathbf{w}_t^k - \overline{\mathbf{w}}^k\| \leq \delta$ for all t and k, and also that $\overline{\mathbf{w}}^k = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t^k$. Therefore $S = \{\mathbf{w}_1^k, \dots, \mathbf{w}_T^k\}$ satisfies the conditions in the infimum in Definition 32 so that $\|\partial F(\overline{\mathbf{w}}^k)\|_\delta \leq \left\|\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\nabla}_t^k\right\|$. □